ANTENNA APERTURE PHASE RETRIEVAL

Peter H. Gardenier

A thesis presented for the degree of
Doctor of Philosophy in Electrical and Electronic Engineering
at the University of Canterbury, Christchurch, New Zealand.

April 1990

ABSTRACT

Geometrical defects of a high gain reflector antenna can cause the radiation pattern of the antenna to fail to meet its specifications. These defects give rise to loss of gain, widening of the main beam and raising of sidelobes. The geometrical defects can be identified, and subsequently corrected, by utilizing information contained in the phase of the copolar aperture field distribution. For technical reasons, this phase can be difficult or inconvenient to measure directly. Therefore, indirect methods of deducing the phase are often preferred.

This thesis introduces an iterative algorithm, called the modified Gerchberg-Saxton algorithm, which has been developed for retrieving the copolar aperture field phase distribution from the far field copolar amplitude pattern. In order to aid convergence of this algorithm, it incorporates information concerning the design and any known aspect of the antenna. The modified Gerchberg-Saxton algorithm is based on the conventional Gerchberg-Saxton algorithm, originally developed for electron microscopy, but incorporates features of Fienup's phase retrieval algorithms.

This thesis reviews radio engineering theory with an emphasis on high gain reflector antennas. In particular, the Fourier transform relationship between the copolar aperture field distribution and the copolar radiation pattern is critically examined. The problem of retrieving the copolar aperture field distribution from the amplitude of its Fourier transform is called a Fourier phase problem. The Fourier phase problem, the uniqueness of its solutions and iterative algorithms for solving it are discussed. Other established methods for determining geometrical defects of an antenna are described and their relative advantages and disadvantages are assessed. The main advantage of the modified Gerchberg-Saxton algorithm is that it requires measurement of only a single copolar amplitude pattern.

The modified Gerchberg-Saxton algorithm is evaluated by applying it to computer simulated data and to measured amplitude patterns of an acoustic antenna. This evaluation illustrates the relationship between the accuracy of the data to which the algorithm is applied and the accuracy of the retrieved copolar aperture field phase distribution. The performance of the algorithm appears to be insensitive to the location and dimensions of the geometrical defects of the antenna. The optimum form of the algorithm seems to be versatile and robust enough to offer real hope of being able to retrieve, to a useful level of accuracy, the phase of the aperture field from a single measured radiation pattern amplitude (i.e. there is no need to measure the phase of the radiation pattern).

ACKNOWLEDGEMENTS

The research for my Ph.D. project and the writing of this thesis could not have come about if it was not for the support of many people. I am especially grateful to my supervisor Professor Richard Bates, for his much valued guidance and encouragement during the course of my research, and for the enormous amount of time and effort he has spent in editing this thesis.

The Telecom Corporation of New Zealand Limited (formally the New Zealand Post Office) has provided me with financial assistance for which I am grateful. I thank Drs Murray Milner and Eric Hamilton, both formally with the International Section of the NZPO Headquarters, for their encouragement and for the helpful discussions we had about practical aspects of setting up and operating earth station satellite antennas.

I am grateful for the cooperation of fellow students, academic staff, technical staff and librarians at the University of Canterbury. I have enjoyed the stimulating discussions with these people and have learned much from them. Special thanks to Charles Parker, Lim Ching Aun, Michael Cusdin, Wiktor Mencel, Quek Bek Kim and Steve Gunn, all of whom have worked with me on aspects of the research described in this thesis.

Thanks to my friends, including my family, flatmates and colleagues, who have encouraged me or lent a listening ear during my time as a post-graduate student. They have been a great source of strength during the stressful time of writing this thesis. I especially thank Heather Kerr for her support and her patience.

CONTENTS

PREFACE

An important way in which engineering science can progress is to take ideas developed for one discipline and to apply them to another discipline. For such an application to suggest itself, the two disciplines must share something in common. My supervisor, Professor R.H.T. Bates, is in a good position to initiate such inter-disciplinary work because he has a diverse range of interests, many of which revolve around the Fourier transform [Bates, 1987a ; 1987b]. His research group in the Department of Electrical and Electronic Engineering at the University of Canterbury has, for the last two decades, been actively researching in areas including general inverse problems (notably computed tomography and ultrasonic imaging), theory and application of image processing, radio antenna engineering and various aspects of biomedical engineering.

I first worked under Professor Bates when I undertook my final year project (for the Bachelor of Electrical and Electronic Engineering degree), which was co-supervised by Professor Bates' research student Alastair Sinton. My project was to investigate a claim by Panarella and Guty [Panarella and Guty, 1983; Panarella, 1985] that optical interference effects reduce at low light levels. If substantial, this claim would have important consequences for radio communications, amongst other things, because the implication appears to be that antenna sidelobes would disappear for very faint signals. The results of experiments that I performed contradicted the claim [Sinton *et al.*, 1986]. This work introduced me to the theory and practice of diffraction and interference effects of electromagnetic waves.

I started my Ph.D. research course, under the supervision of Professor Bates, by working on two projects. The first project involved the problem of determining the depolarization of a high gain reflector antenna from measurements made with a source antenna which itself suffers from an unknown amount of depolarization. A standard method for solving this problem involves physically rotating the high gain reflector antenna, or its feed, by 90° about its axis. It is preferable, however, to be able to dispense with such physical manipulation of the antenna. It was felt that some kind of holographic-type approach, such as had earlier been applied to the radio engineering phase problem (Sec. 3.5.2), might be helpful. However, after studying the problem for a while, no solution suggested itself to me.

The second project was concerned with an aspect of pattern recognition. I worked on this project with Bruce McCallum, a fellow research student. Because of their invariances to various symmetry operations, the amplitudes of Fourier and related transforms are often used in pattern recognition contexts. Reitboeck and Altmann [1984] reason that, when the Fourier phase information is lost, there is likely to be a very large number of patterns of different shapes whose Fourier transform amplitudes would be identical. Professor Bates' wide experience of phase problems convinced him that this is not true in general. We therefore devised our own descriptor of objects, based upon their Fourier transform amplitudes. This descriptor is insensitive to the location, ori-

entation, magnification and brightness of the object [McCallum *et al.*, 1986] and,in general, is also uniquely invertible [Gardenier *et al.*, 1986a].

While I was working on the above-mentioned projects, Professor Bates and one of his research students, David Tan, were supervising the final year project of Lim Ching Aun. This project was a preliminary study of the potential usefulness for the radio engineering phase problem of the Gerchberg-Saxton algorithm (Sec. 3.4.3.1), which was originally developed for electron microscopy. I continued this work and my development of it forms the basis of this thesis. Lim Ching Aun later completed a Master of Engineering degree for which he studied how the Gerchberg-Saxton algorithm (that had by then been modified by myself) can be applied to determine the causes of depolarization of a high gain reflector antenna (assuming that the source antenna does not itself suffer from depolarization). An adapted version of this work is reported in Section 4.9.1.

The radio engineering phase problem, mentioned in the previous paragraph, involves determining an antenna's copolar aperture field distribution when given only the amplitude of its copolar radiation pattern. The radio engineering phase problem often arises for high gain reflector antennas whose copolar phase patterns are difficult to measure. The reason for wanting to determine the copolar aperture field distribution is that its phase provides valuable information about those geometrical defects of the antenna which must be corrected before the antenna can perform optimally. I have developed a modified form of the Gerchberg-Saxton algorithm suitable for solving the radio engineering phase problem. In this thesis it is demonstrated, on the basis of results of computer simulations and an experiment with an acoustic antenna, that this modified Gerchberg-Saxton algorithm provides a potentially practicable method of retrieving the copolar aperture field phase distribution from a single measured amplitude pattern (hence the title of this thesis: antenna aperture phase retrieval).

This thesis is written in six chapters. Each chapter concludes with a summary of the main points raised in that chapter. A chapter by chapter outline of the thesis now follows.

Chapter 1 provides an overview of antenna engineering. The fundamentals of electromagnetic wave theory are introduced and the main types of antenna are briefly described. The chapter also discusses the electrical properties of antennas and the ways in which radio waves propagate through the earth's atmosphere.

Chapter 2 concentrates on high gain reflector antennas. Different methods for analysing the scattering from reflectors are introduced. This leads to the derivation of the Fourier transform relationship between the copolar aperture field distribution and the copolar far field pattern. This relationship is of central importance for the modified Gerchberg-Saxton algorithm. Different configurations of reflector antennas are described. Relevant characteristics of the radiation pattern of a high gain antenna are discussed, with particular emphasis on the degradation of the radiation pattern that occurs when the copolar aperture field phase distribution is distorted by geometrical defects of the antenna. The chapter also mentions typical applications for high gain reflector antennas.

Chapter 3 provides background for the development of the modified Gerchberg-Saxton algorithm. The way that different geometrical defects arise are discussed. It is shown that many geometrical defects cause deviations in the copolar aperture field phase distribution and can be inferred from these deviations. Various methods for deducing the copolar aperture field phase distribution are discussed. It is argued that, in many situations, it is desirable to be able to deduce the copolar aperture field phase distribution from only the amplitude of the copolar radiation pattern. Thus it is nec-

essary to solve the radio engineering phase problem (defined earlier in this preface). The radio engineering phase problem is a specialization of the Fourier phase problem. The Fourier phase problem is discussed in detail and it is shown that, in general, it has a unique solution in two dimensions. Existing ways of solving both the Fourier phase problem and the radio engineering phase problem are outlined.

In Chapter 4 the modified Gerchberg-Saxton algorithm is developed. The chapter starts by suggesting ways in which the algorithm could be applied in practice. A generalized computer model of a high gain reflector antenna and the measurement process is then defined. This model is invoked to generate a wide variety of data to which the modified Gerchberg-Saxton algorithm can be applied. The results of applying the algorithm to many different computer generate data are presented, evaluated and discussed.

In Chapter 5 the modified Gerchberg-Saxton algorithm is applied to data obtained by measuring the amplitude pattern of a sonic antenna. It is shown that the radiation pattern and the aperture field distribution of a sonic antenna are related by the Fourier transform relationship. The apparatus with which the amplitude pattern is measured is described. This chapter also argues that the modified Gerchberg-Saxton algorithm can be usefully applied to data other than those obtained from high gain reflector antenna patterns.

In Chapter 6 a number of avenues for future research are discussed. This chapter and the thesis concludes with a summary of the main features of the modified Gerchberg-Saxton algorithm.

My original research constitutes most of the material described in Chapters 4 and 5. Almost all of the software implementing the modified Gerchberg-Saxton algorithm and the computer model was written by me. This software was written to interface to the **improc** (**image processing**) utility which was principally developed by Richard Lane, who is another of Professor Bates' former research students. Most of the electronics for the acoustic experiment were built by Wiktor Mencel, under the supervision of Michael Cusdin. My fellow research student Charles Parker's first project, under Professor Bates' supervision, was to assist me with the measurements. He perfected the measurement technique, which he has now incorporated into his own research programme.

Papers and presentations prepared during the course of my Ph.D. research are listed below.

SINTON, A.M., GARDENIER, P.H. and BATES, R.H.T. (1986), 'Reinvestigation of optical interference at low light levels', *Speculations in Science and Technology*, Vol. 9, No. 4, November, pp. 269–278.

McCALLUM, B.C., GARDENIER, P.H. and BATES, R.H.T. (1986), 'Invertible invariant transformations for robotic catalogues', in *Proceedings of the International Conference on Future Computing Systems*, Christchurch, New Zealand, February, pp. 151–158.

GARDENIER, P.H., McCALLUM, B.C. and BATES, R.H.T. (1986a), 'Fourier transform magnitudes are unique pattern recognition templates', *Biological Cybernetics*, Vol. 54, No. 6, September, pp. 385–391.

GARDENIER, P.H., LIM, C.A., TAN, D.G.H. and BATES, R.H.T. (1986b), 'Aperture distribution phase from single radiation pattern measurement via Gerchberg-Saxton algorithm', *Electronics Letters*, Vol. 22, No. 2, January, pp. 113–115.

GARDENIER, P.H., LIM, C.A., TAN, D.G.H. and BATES, R.H.T. (1986c), 'Feed-position and reflector-shape errors of satellite communications antenna from radiation pattern magnitude', in *IPENZ Conference 86*, Auckland University, New Zealand, February.

BATES, R.H.T., FRIGHT, W.R. and GARDENIER, P.H. (1987), 'Gerchberg-Saxton phase retrieval when image magnitude given only approximately', in IDELL, P.S. (Ed.), *Digital Image Recovery and Synthesis*, Proceedings of the SPIE Volume 828, August, pp. 171–176.

MILNER, M.O., GARDENIER, P.H. and BATES, R.H.T. (1987), 'Antenna aperture phase from far field magnitude', in *IEEE/AP-S International Symposium and URSI Radio Science Meeting*, Virginia Tech, Blacksburg, Virginia, USA, June.

GARDENIER, P.H., LIM, C.A. and PARKER, C.R. (1988), 'Satellite communications antenna misalignments inferred from far field magnitude', in *Proceedings of the 25th New Zealand National Electronics Conference*, NELCON, Christchurch, August, pp. 83–88.

CHAPTER 1

OVERVIEW OF ANTENNA ENGINEERING

The existence of radio waves was confirmed just over 100 years ago by Hertz [O'Hara and Pricha, 1987]. Since then, these waves have become an integral part of modern life. The world has been transformed by radio, and even more by television.

This chapter provides a brief overview of antenna engineering. Section 1.1 introduces the theory of electromagnetic waves (of which radio waves are a subset), starting with Maxwell's equations. Radio waves are transmitted and received by antennas. The properties and different types of antennas are summarized in Sections 1.2 and 1.3. The effects on radio wave propagation of the earth and its atmosphere are discussed in Section 1.4. This chapter concludes with a summary of important practical applications of antennas and radio waves.

1.1 ELECTROMAGNETIC WAVES

The electromagnetic spectrum covers a broad range of frequencies, from near zero cycles/second (Hz), to gamma ray frequencies (Fig. 1.1). *Radio waves* are electromagnetic waves which have a frequency such that they may be detected and amplified as an electric current of the same frequency [IEEE, 1984]. Radio frequencies are at present limited, by technological constraints, to the range from about 10 kHz to about 100 GHz — the region between audio frequencies and infra-red frequencies. *Microwaves* (also referred to as *short waves*) are loosely defined as radio waves with a frequency of 1 GHz and higher (and therefore a wavelength of 0.3 m or shorter), while radio waves with a lower frequency (and longer wavelength) are called *long waves* [Silver, 1949, p. 2].

1.1.1 Harmonically time varying fields

An electromagnetic wave consists of time varying electric and magnetic fields. It is appropriate to consider harmonically time varying fields because:

1. Any time varying quantity can be expressed as the summation of harmonic components.

2. In practice, many generators produce electric and magnetic fields which are approximately harmonic.

3. Analysis of frequency dependent systems can be simplified by examining their behaviour within a succession of contiguous bands, each of which is narrow enough to be treated as if it effectively single frequency.

A vector, for example the electric field intensity vector $\mathbf{E}$, which varies with both spatial position $\mathbf{r}$ and time t, can be expressed as a function of these two quantities

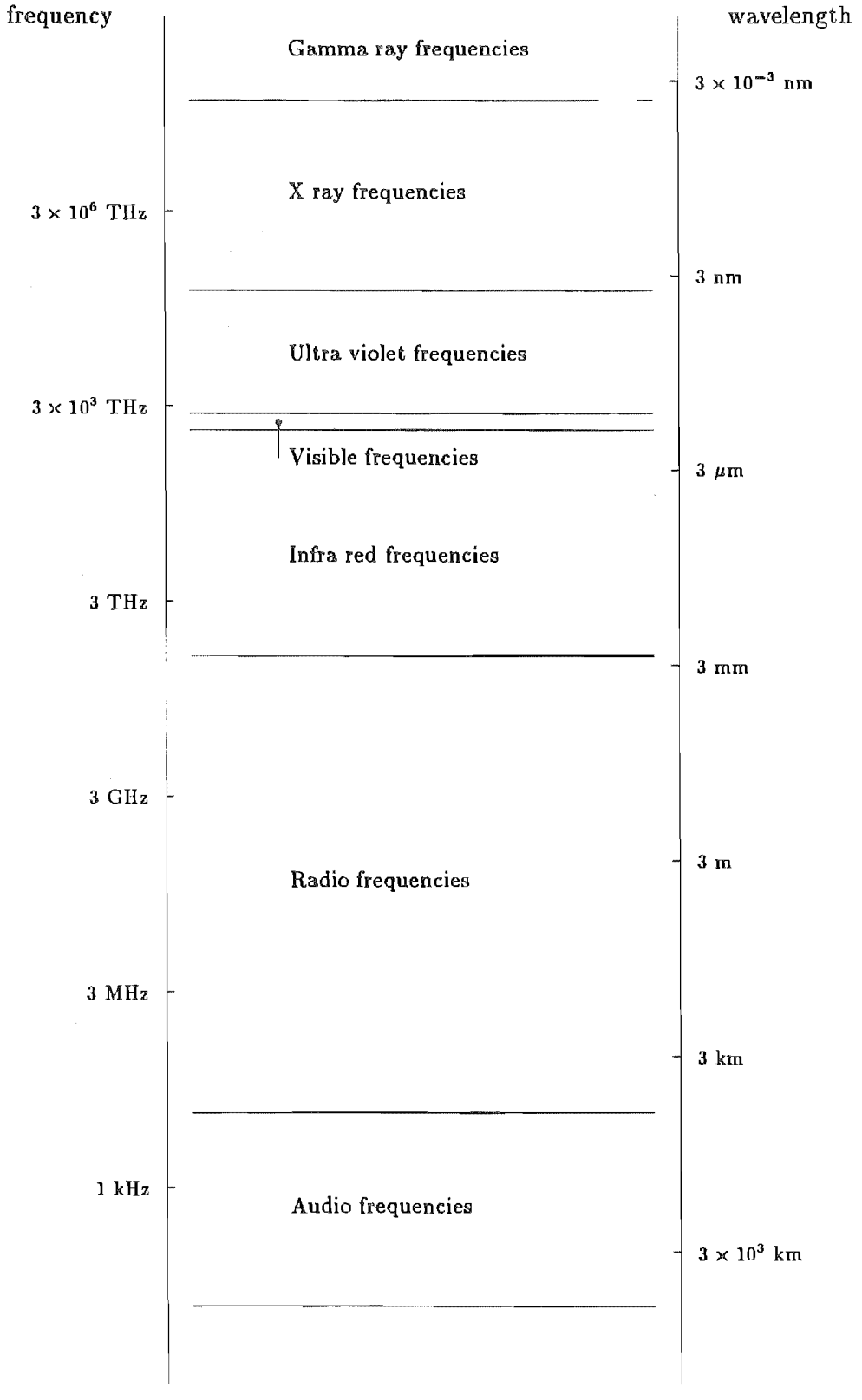


Figure 1.1 Some of the frequency bands of the electromagnetic spectrum (frequency boundaries are approximately those given by IEEE [1984]).

as $\mathbf{E}(\mathbf{r},t)$. When it varies harmonically with time at each point in space, it can be expressed as

$$\mathbf{E}(\mathbf{r},t) = \text{real}\left\{\mathbf{E}(\mathbf{r})e^{j\omega t}\right\} \tag{1.1}$$

where $\mathbf{E}(\mathbf{r})$ is a complex vector function of spatial position but not of time and where ω is angular frequency. Throughout this thesis, vector quantities are denoted by boldface letters. Time varying (real) quantities are expressed as static complex quantities with the time dependence $e^{j\omega t}$ omitted (but nevertheless understood).

The *vector component* of $\mathbf{E}$ in the direction of a unit vector $\hat{\mathbf{x}}$ is identified by a subscript 'x' and is defined by

$$E_x = \mathbf{E}\cdot\hat{\mathbf{x}} \tag{1.2}$$

where the dot denotes the inner scalar product operation. The x component of $\mathbf{E}$ is a complex scalar, which is expressed in terms of its *amplitude* and *phase* as

$$E_x = |E_x|e^{\text{phase}\{E_x\}} \tag{1.3}$$

where the $|\cdot|$ notation for a scalar quantity denotes its amplitude. Note that

$$|E_x| = (E_x E_x^*)^{1/2} \tag{1.4}$$

where the asterisk denotes the complex conjugate of a scalar.

The complex vector $\mathbf{E}$ can be expressed in terms of its Cartesian components in the following way:

$$\mathbf{E} = E_x\,\hat{\mathbf{x}} + E_y\,\hat{\mathbf{y}} + E_z\,\hat{\mathbf{z}} \tag{1.5}$$

The *magnitude* of a vector is denoted by $|\cdot|$ and defined to be

$$|\mathbf{E}| = (\mathbf{E}\cdot\mathbf{E}^*)^{1/2} = \left(|E_x|^2 + |E_y|^2 + |E_z|^2\right)^{1/2} \tag{1.6}$$

The magnitude of a real vector is equivalent to its length.

The *Cartesian coordinates* of $\mathbf{r}$ are written as (x,y,z) where

$$x = \mathbf{r}\cdot\hat{\mathbf{x}}\,,\quad y = \mathbf{r}\cdot\hat{\mathbf{y}}\,,\quad z = \mathbf{r}\cdot\hat{\mathbf{z}} \tag{1.7}$$

Points in space can also be described by the spherical coordinate system defined in Figure 1.2. The *spherical coordinates* of $\mathbf{r}$ are denoted by $(r;\theta;\phi)$, where semicolons inside parenthesis always delimit spherical ordinates [Bates and McDonnell, 1989, Sec. 6]. The spherical ordinates are related to Cartesian ordinates by

$$\begin{aligned} x &= r\sin\theta\cos\phi \\ y &= r\sin\theta\sin\phi \\ z &= r\cos\theta \end{aligned} \tag{1.8}$$

The notation can be extended to describe the spatial variation of, say $\mathbf{E}$, in the following different ways: $\mathbf{E}(\mathbf{r}) = \mathbf{E}(x,y,z) = \mathbf{E}(r;\theta;\phi)$.

The definitions and notation introduced in this section apply to any vectors (such as field intensity or position vectors) and to any set of orthogonal unit vectors (not only Cartesian unit vectors).

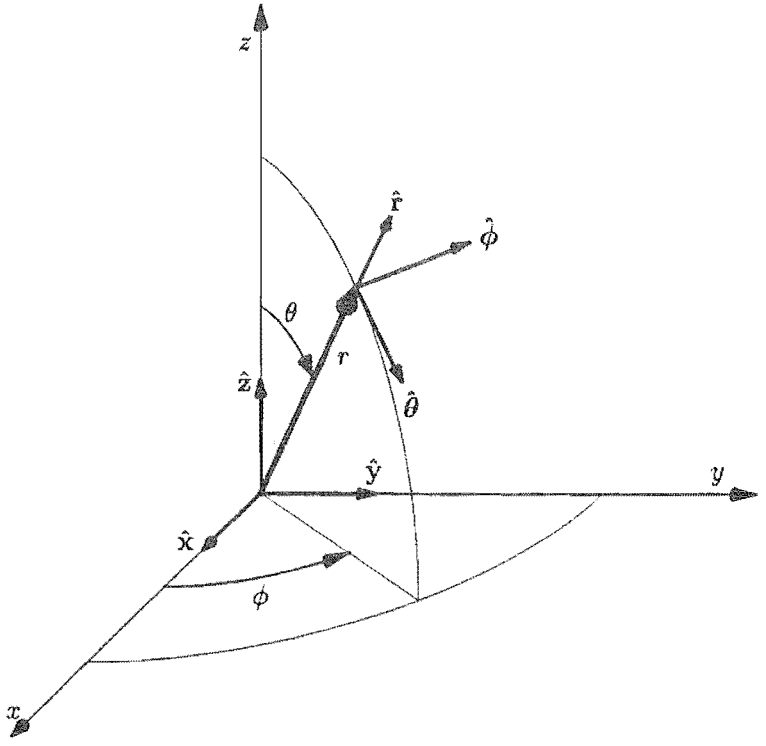


Figure 1.2 Graphical representation of the spherical coordinate system in relation to the Cartesian coordinate system.

1.1.2 Maxwell's equations

Equations describing electromagnetic waves were formulated by Maxwell [1865]. The time harmonic versions of his equations are [cf. Rudge *et al.*, 1982, p. 7; Jordan and Balmain, 1968, Chap. 4]

$$\begin{aligned} \nabla\times\mathbf{E} &= -j\omega\mu\mathbf{H} - \mathbf{J_m} \\ \nabla\times\mathbf{H} &= (\sigma + j\omega\epsilon)\mathbf{E} + \mathbf{J} \\ \mu\nabla\cdot\mathbf{H} &= \rho_\mathbf{m} \\ \epsilon\nabla\cdot\mathbf{E} &= \rho \end{aligned} \tag{1.9}$$

where $\mathbf{E}$ and $\mathbf{H}$ are the *electric field* and *magnetic field* respectively, and

$$\begin{aligned} \mathbf{J} &= \text{the electric current density,} \\ \mathbf{J_m} &= \text{a fictitious magnetic current density,} \\ \rho &= \text{the electric charge density,} \\ \rho_\mathbf{m} &= \text{a fictitious magnetic charge density,} \\ \sigma &= \text{the conductivity of the medium,} \\ \epsilon &= \text{the electric permittivity of the medium,} \end{aligned} \tag{1.10}$$

$$\mu = \text{the magnetic permeability of the medium,}$$

$$\nabla = \text{the vector differential operator}$$
$$= \frac{\partial}{\partial x}\hat{\mathbf{x}} + \frac{\partial}{\partial y}\hat{\mathbf{y}} + \frac{\partial}{\partial z}\hat{\mathbf{z}}$$

The quantities $\mathbf{J}$ and ρ are considered to be the *sources* of the electromagnetic wave. The magnetic sources $\mathbf{J}_m$ and ρ_m are fictitious quantities which are often mathematically convenient to invoke (as in, for example, Sec. 2.1.3.1), although from a strictly factual, physical point of view, it is necessary to set $\mathbf{J}_m = \rho_m = 0$. The sources vary spatially and time harmonically.

The quantities σ, ϵ and μ, are called the *constitutive parameters* of the medium and are functions of spatial position. In general they can also be functions of time, but in the time harmonic formulation (1.9) they are assumed to be temporally constant. The constitutive parameters often vary with angular frequency, which in (1.9) is represented by ω.

1.1.3 Types of media

A medium is *homogeneous* if the constitutive parameters are constant throughout it. An *isotropic* medium is one in which the constitutive parameters are scalars (implying that they do not depend on the direction of the electric and magnetic fields).

A *linear* medium is one in which the constitutive parameters do not vary with intensity of the electric and magnetic fields. In a linear medium the *principle of superposition* holds. This states that if a source distribution produces electric and magnetic fields $(\mathbf{E}_1, \mathbf{H}_1)$, and another source distribution produces $(\mathbf{E}_2, \mathbf{H}_2)$, then when the two source distributions are applied together, the resultant fields are $(\mathbf{E}_1 + \mathbf{E}_2, \mathbf{H}_1 + \mathbf{H}_2)$.

A *good dielectric* is a good insulator, implying that $\sigma \ll \omega\epsilon$ (compare with the term in braces in the second equation of (1.9)). A *perfect dielectric* is non-conducting, implying that $\sigma = 0$. A *lossless* medium is a perfect dielectric in which both ϵ and μ are real (that is, have no imaginary part).

Free space is a vacuum and is therefore a linear, homogeneous, isotropic, perfect dielectric containing no sources. Free space has a permeability, denoted by μ_v, defined to be $4\pi \times 10^{-7}$ henry/metre, and a permittivity, denoted by ϵ_v, which is very close to $(1/36\pi) \times 10^{-9}$ farad/metre.

A *good conductor* is a medium in which $\sigma \gg \omega\epsilon$, while for a *perfect conductor* $\sigma = \infty$. From the second equation of (1.9) it can be seen that, for $\nabla \times \mathbf{H}$ and $\mathbf{J}$ finite, $\mathbf{E}$ must tend to zero as σ tends to infinity. The first equation of (1.9) shows that $\mathbf{H}$ tends to zero as $\mathbf{E}$ tends to zero (since $\mathbf{J}_m = 0$). Therefore, an electromagnetic field cannot exist inside a perfect conductor.

1.1.4 Boundary conditions

Equations (1.9) are the derivative form of Maxwell's equations and are only applicable within a continuous medium. To express the behaviour of an electromagnetic wave at points of discontinuity of any of the constitutive parameters, the integral form of Maxwell's equations [Jordan and Balmain, 1968, Sec. 4.04] can be applied.

Consider two media separated by a boundary surface (see Fig. 1.3). The subscripts '1' and '2', which refer to medium 1 and 2 respectively, are used for quantities at

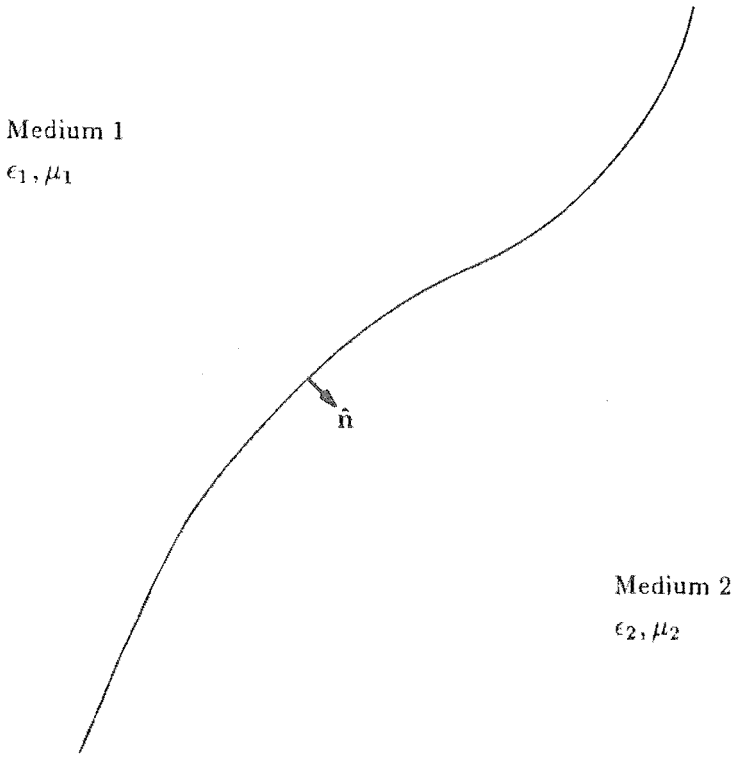


Figure 1.3 A surface forming the boundary between two media.

contiguous points on either side of the boundary. Application of the integral equations at the boundary surface yields *boundary conditions* [Silver, 1949, Sec. 3.3]

$$\begin{aligned} \hat{\mathbf{n}} \times (\mathbf{E}_2 - \mathbf{E}_1) &= -\mathbf{J}_{ms} \\ \hat{\mathbf{n}} \cdot (\epsilon_2 \mathbf{E}_2 - \epsilon_1 \mathbf{E}_1) &= \rho_s \\ \hat{\mathbf{n}} \times (\mathbf{H}_2 - \mathbf{H}_1) &= \mathbf{J}_s \\ \hat{\mathbf{n}} \cdot (\mu_2 \mathbf{H}_2 - \mu_1 \mathbf{H1}) &= \rho_{ms} \end{aligned} \tag{1.11}$$

where $\hat{\mathbf{n}}$ is the unit normal to the surface, pointing from medium 1 to medium 2. The terms on the right hand side of (1.11) are called *surface currents* and *surface charges* and are distributed over the boundary surface. They are differentiated from the volume sources of (1.9) by the subscript 's'.

An important situation arises when one of the media, say medium 1, is a perfect conductor. As pointed out in Section 1.1.3, both $\mathbf{E}_1$ and $\mathbf{H}_1$ are identically zero, so that (1.11) simplifies to

$$\begin{aligned} \hat{\mathbf{n}} \times \mathbf{E}_2 &= 0 \\ \hat{\mathbf{n}} \cdot \epsilon_2 \mathbf{E}_2 &= \rho_s \\ \hat{\mathbf{n}} \times \mathbf{H}_2 &= \mathbf{J}_s \\ \hat{\mathbf{n}} \cdot \mu_2 \mathbf{H}_2 &= 0 \end{aligned} \tag{1.12}$$

1.1.5 Wave equations

An electromagnetic wave is fully characterized by either its electric field $\mathbf{E}$ or its magnetic field $\mathbf{H}$. By convention, and throughout this thesis, $\mathbf{E}$ is usually invoked. From it, one can always calculate $\mathbf{H}$ (provided the properties of the medium are known) by rearranging the first equation of (1.9):

$$\mathbf{H} = \frac{j}{\omega\mu}\nabla\times\mathbf{E} \tag{1.13}$$

For a homogeneous, linear medium containing sources, Maxwell's equations (1.9) can be transformed into *wave equations* (also called vector Helmholtz equations [Silver, 1949, Sec. 3.6]):

$$\begin{aligned} \nabla^2\mathbf{E} + \gamma^2\mathbf{E} &= j\omega\mu\mathbf{J} + \nabla\times\mathbf{J}_\mathbf{m} + \frac{1}{\epsilon}\nabla\rho \\ \nabla^2\mathbf{H} + \gamma^2\mathbf{H} &= j\omega\epsilon\mathbf{J}_\mathbf{m} - \nabla\times\mathbf{J} + \frac{1}{\mu}\nabla\rho_\mathbf{m} \end{aligned} \tag{1.14}$$

where $\gamma = [(-j\omega\mu)(\sigma + j\omega\epsilon)]^{1/2}$ is called the *propagation constant.* For lossless media, it is real and equals the *wave number* [Ramo *et al.*, 1965, p. 326]:

$$k = \frac{2\pi}{\lambda} = \omega(\mu\epsilon)^{1/2} \tag{1.15}$$

where λ is the *wavelength* of the wave. The speed of propagation of the wave is $v = \omega/\gamma$. For free space, the speed of propagation reduces to $c = (\mu_\mathrm{v}\epsilon_\mathrm{v})^{-1/2}$ which is close to 3×10^8 metres/second. The *index of refraction* of any medium is defined as $n = c/v$. The attenuation suffered by a wave as it propagates through a lossy medium is determined by the imaginary part of γ [Jordan and Balmain, 1968, p. 126]. In a source free medium, all the terms on the right hand sides of (1.14) are zero.

The *Complex Poynting vector* $\mathbf{P}$ is defined by [ITT, 1968, p. 43-3]

$$\mathbf{P} = \frac{1}{2}(\mathbf{E}\times\mathbf{H}^*) \tag{1.16}$$

It follows from Maxwell's equations and the law of conservation of energy, that the integral of the normal component of $\mathbf{P}$, over the surface enclosing any volume, is the total power flowing out of that volume. Poynting's theorem states that, at each point in space, $\mathbf{P}$ gives the magnitude and direction of the average power flow (i.e. power per unit area) [Jordan and Balmain, 1968, Chap. 6].

A *ray* is a curve through space, such that at each point along its length, it is parallel to $\mathbf{P}$ [Silver, 1949, p. 110]. A *pencil of rays* is a thin bent rod shaped volume, of variable cross-section, surrounding a central ray and bounded by a family of rays lying on its surface. From energy conservation principles [Born and Wolf, 1970, p. 115], the total power flow through any cross-section of a given pencil is constant in a lossless medium. Therefore, the intensity of a field along a ray is inversely proportional to the cross-sectional area of a surrounding pencil of rays.

1.1.6 Polarization

The *polarization* of an electric field vector, at each point in space, is the locus of the extremity of its real part (see (1.1)), when the magnitude of the real vector is envisaged

as a distance from the point in space. For a fixed frequency, the locus is an ellipse which lies in the *polarization plane* [IEEE, 1984].

A field is *linearly polarized* when the minor axis of the ellipse vanishes, implying that the electric field vector is at all times pointing in the same direction. Two special cases of linear polarization are *horizontal polarization*, in which the electric field vector is parallel to the earth's surface, and *vertical polarization*, in which the electric field vector is perpendicular to the earth's surface.

A field is *circularly polarized* when the ellipse degenerates into a circle, implying that the magnitude of the electric field vector is always constant. A field which is neither linearly polarized nor circularly polarized is called an *elliptically polarized* field. The *sense of polarization* is the sense in which the ellipse is traversed by the extremity of the real part of the electric field vector.

The *polarization unit vector* of a field is defined to be [IEEE, 1984]

$$\hat{\imath} = \frac{\mathbf{E}}{|\mathbf{E}|} \tag{1.17}$$

The polarization unit vector of a field completely describes the polarization of the field.

Two polarization unit vectors $\hat{\imath}_1$ and $\hat{\imath}_2$, lying in the polarization plane of a field vector $\mathbf{E}$, are *orthogonal* if $\hat{\imath}_1 \cdot \hat{\imath}_2^* = 0$. The vector $\mathbf{E}$ is completely described by its (orthogonal) vector components E_1 and E_2 where (cf. (1.2) and (1.5))

$$E_1 = \mathbf{E} \cdot \hat{\imath}_1^* \qquad \text{and} \qquad E_2 = \mathbf{E} \cdot \hat{\imath}_2^* \tag{1.18}$$

When the Cartesian x and y axes lie in the polarization plane, $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ form an orthogonal pair of linearly polarized unit vectors. Another possible pair of orthogonal unit vectors are left and right hand circularly polarized [Rumsey *et al.*, 1951, part III]:

$$\hat{\imath}_{\mathrm{R}} = \frac{1}{\sqrt{2}}(\hat{\mathbf{x}} - j\,\hat{\mathbf{y}}), \qquad \hat{\imath}_{\mathrm{L}} = \frac{1}{\sqrt{2}}(\hat{\mathbf{x}} + j\,\hat{\mathbf{y}}) \tag{1.19}$$

Note that there can be no component of $\mathbf{E}$ in the direction perpendicular to the polarization plane.

For arbitrarily oriented Cartesian axes, a field vector is completely defined by the amplitude and phase of all Cartesian components (cf. Sec. 1.1.1). Only the phase and relative amplitude of each component is required to define a field's polarization. The polarization of a field is a function of spatial position.

1.2 ELECTRICAL PROPERTIES OF ANTENNAS

An antenna is defined as "a means of radiating or receiving radio waves" [IEEE, 1984]. As seen from the wave equations (1.14), the sources of electromagnetic radiation are charges and currents, where a current is composed of moving charges. A *transmitting antenna* provides suitable conditions for radio frequency electric currents to radiate an electromagnetic wave. A *receiving antenna* intercepts a radio wave, which induces electric currents on the antenna.

Parameters which describe the electrical performance of an antenna include: antenna pattern, gain and efficiency, polarization, impedance, frequency of operation, bandwidth and noise temperature. An antenna is designed to meet specifications on some or all of these electrical parameters, as well as meeting mechanical, environmental and cost constraints. The electrical properties of antennas, including the performance parameters, are discussed in the following sections.

1.2.1 Field regions

The electromagnetic wave radiated from an antenna can be considered to be composed of two fields. The *radiating field* is that part of the wave which transports energy away from the antenna. The *reactive field* oscillates energy between the space near to the antenna and the antenna itself. It is useful to divide the space surrounding the antenna into regions which are differentiated by different characteristics of the electromagnetic wave [Rudge *et al.*, 1982, Sec. 1.4].

The *reactive near field region* is the region close to the antenna, where the reactive field dominates the radiating field. The reactive field decays faster than the radiating field and, for most antennas, the outer limit of the reactive near field region is of the order of a few wavelengths or less.

The *far field region* is far enough from the antenna that the angular distribution of the field is independent of distance [IEEE, 1979, p. 139]. The wave consists effectively of only the radiating field, which decays with the inverse of distance from the antenna. The far field region is sufficiently distant that the relative contributions to the radiating field from different parts of the antenna are independent of distance. This implies that the antenna can be treated as if it were a point source with directional variations. Although this condition is only exact infinitely far from the antenna, it is adequately approximated at finite distances greater than the following lower bound [Blake, 1984, p. 122]:

$$R_{\mathrm{ff}} = \frac{2D^2}{\lambda} \tag{1.20}$$

where R_{ff} is the accepted *minimum far field distance* from an antenna whose largest dimension is D. For small antennas, R_{ff} should be no less than a wavelength. The part of the radiating field which is in the far field region is called the *far field.*

The *radiating near field region* occupies the space between the reactive near field region and the far field region. The radiating field dominates the reactive field, but the relative angular distribution of the radiating field depends on distance from the antenna. The radiating near field region is close enough that the size of the antenna is significant, so the relative contributions to the radiating field from different parts of the antenna depend on distance as well as on angle. The part of the radiating field which is in the near field region is called the *near field.* The radiating near field region does not exist for antennas which are small compared to a wavelength.

1.2.2 Antenna patterns

An *antenna pattern* is the spatial distribution of a quantity which characterizes the electromagnetic field generated by an antenna [IEEE, 1984]. The quantity is usually determined over the surface of a sphere which is centred on the antenna. A spherical coordinate system (Sec. 1.1.1), whose origin is at the centre of the sphere, is utilized to locate points on the surface. Only the two angular ordinates $(\theta; \phi)$ are required to specify a point on the sphere, since the radius ordinate is constant.

The *gain pattern* $G(\theta; \phi)$ of a transmitting antenna is

$$G(\theta; \phi) = \frac{4\pi P_{\mathrm{rad}}(\theta; \phi)}{P_{\mathrm{in}}} \tag{1.21}$$

where P_{rad} is the power radiated per unit solid angle in direction $(\theta; \phi)$ and is determined in the far field region. P_{in} is the total power accepted from the source. The gain defines the ability of an antenna to concentrate power in a particular direction and it accounts

for power losses within the antenna. The *peak gain* $G_{\max}$ of an antenna is the maximum value of its gain pattern. An antenna with a high peak gain is said to be more directional than an antenna with a lower peak gain and the same power losses.

The *effective area* A_e of a receiving antenna, connected to a matched load, in the direction $(\theta;\phi)$ is

$$A_e(\theta;\phi) = \frac{P_{\text{out}}}{P_{\text{inc}}(\theta;\phi)} \tag{1.22}$$

where P_{out} is the power delivered to the load and $P_{\text{inc}}(\theta;\phi)$ is the power per unit area of an incident wave radiated from a distant source located at angle $(\theta;\phi)$. The received wave is polarized to produce maximum power output from the antenna (see Sec. 2.2.6). The effective area is a measure of how much power can be transferred from an incident wave to the load.

The *amplitude pattern* of an antenna is the angular distribution of the amplitude, or the relative amplitude, of a vector component of the electric field. The *phase pattern* of an antenna is the angular distribution of the phase (relative to some reference phase) of a vector component of the electric field. The *polarization pattern* of an antenna is the angular distribution of the polarization (Sec. 1.1.6) of the electric field.

The *radiation pattern* of an antenna is the angular distribution of the complex electric field vector. When the radiation pattern is determined in the far field region it can be called the *far field pattern*. The far field pattern is is completely characterized by the amplitude and phase patterns of all (complex) vector components of the far field, the peak gain and the power fed to the antenna.

1.2.3 Reciprocity

The *principle of reciprocity* can be invoked to relate the transmitting characteristics of an antenna to its receiving characteristics.

Consider the measurement of the transmitting characteristics of antenna A, taken at an angle θ (Fig. 1.4(a)). In practice this is done with the aid of a second antenna B which is sufficiently far from A to avoid multiple interactions. A voltage source connected to the terminals of A produces a current i_1 through the load connected to the terminals of B.

Now consider the measurement of the receiving characteristics of A, made for the same angle θ (Fig. 1.4(b)), by swapping the voltage source and the load, without moving either antenna. The same voltage source (now connected to the terminals of B) produces a load current i_2 at A.

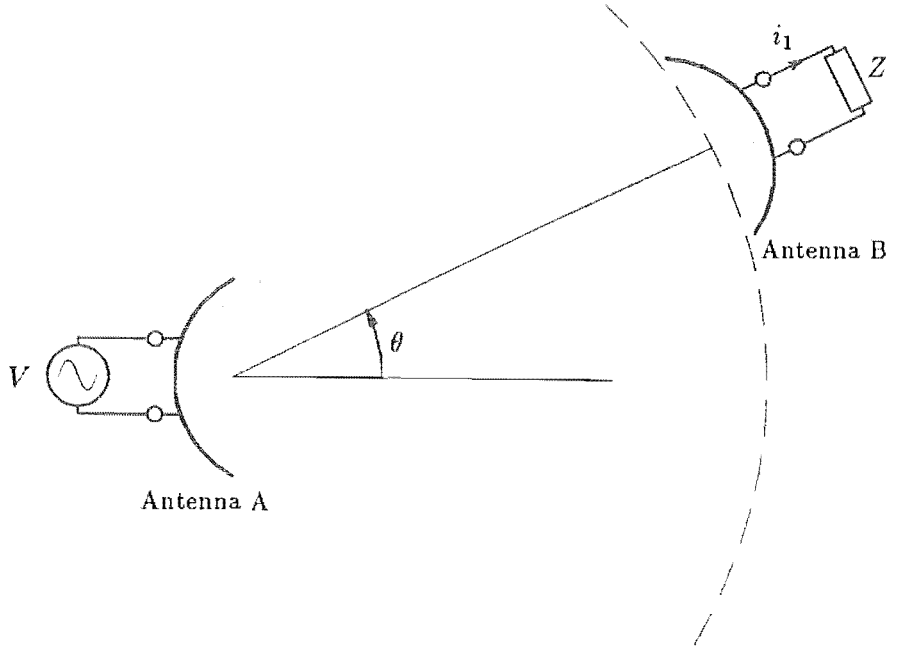
The principle of reciprocity says that the i_2/i_1 is a constant independent of θ, provided that the antenna system is reciprocal [IEEE, 1979, p. 142]. An antenna system with no ferrite or plasma devices, and embedded in a linear isotropic transmission medium, is reciprocal.

Therefore, the directional properties of an antenna which is transmitting are proportional to the directional properties of the same antenna when it is receiving. This means that, in practice, the radiation pattern of an antenna can be measured while it is either transmitting or receiving.

It follows from the reciprocity principle that the effective area of an antenna is related to its gain pattern by the relation [Collin and Zucker, 1969a, Sec. 4.4]

$$A_e(\theta;\phi) = \frac{\lambda^2}{4\pi} G(\theta;\phi) \tag{1.23}$$

(a)



(b)

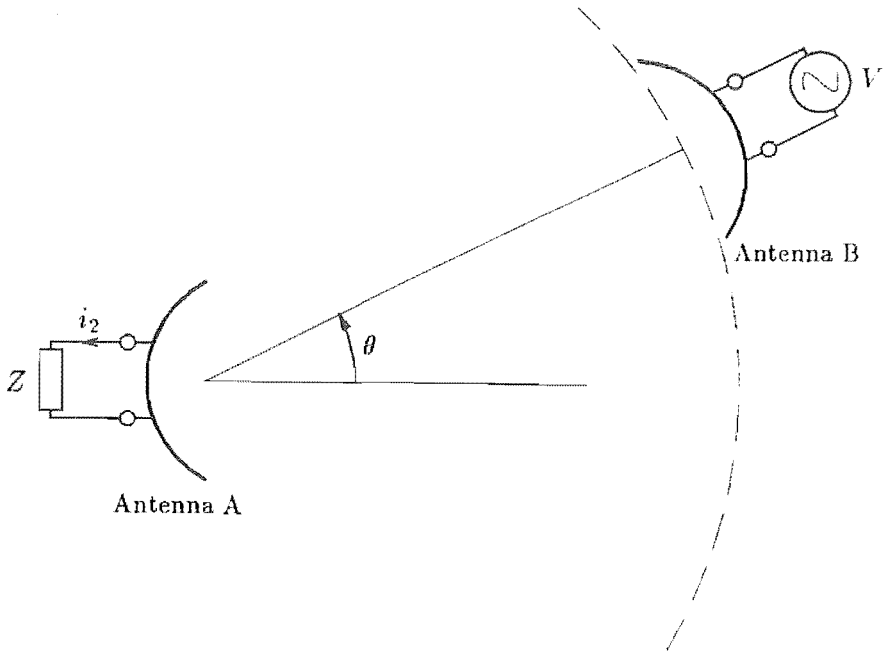


Figure 1.4 Arrangement of two antennas to demonstrate reciprocity. The subfigures are for antenna A (a) transmitting and (b) receiving.

1.2.4 Impedance

In practice, an antenna is connected to an electrical circuit. When transmitting, the circuit provides the current to drive the antenna, and when receiving, the circuit acts as an electrical load for the antenna, abstracting the information carried by the radio wave. In either case, the electrical circuit sees the antenna as an impedance, called the *antenna impedance*. To transfer maximum power to and from the antenna, the antenna impedance must be matched to the impedance (seen by the antenna) of the circuit.

The imaginary part of the impedance for a transmitting antenna is due to the reactive energy which is stored in the reactive field of the antenna. The real part of the antenna resistance is made up of a *loss resistance* R_{loss} in series with the *radiation resistance* R_{rad}. The loss resistance manifests ohmic and dissipative losses in the antenna. The radiation resistance has a value equal to that of the equivalent resistor which would dissipate the same amount of power that the antenna radiates, if it was connected to the circuit in place of the antenna.

The *radiation efficiency* η_{rad} of an antenna is the ratio of the power radiated by the antenna to the total input power [Rudge *et al.*, 1982, Sec. 1.6]. Because the power dissipation in each of two resistors in series is proportional to the respective values of these resistors, it follows that

$$\eta_{\text{rad}} = \frac{R_{\text{rad}}}{(R_{\text{loss}} + R_{\text{rad}})} \tag{1.24}$$

Thus an antenna is efficient if the radiation resistance is much greater than the loss resistance.

1.2.5 Frequency of operation

The antenna performance parameters described so far are defined for a given fixed frequency. However antennas must usually operate over a range of frequencies.

The *bandwidth* of an antenna is defined as the frequency range over which the antenna meets all of its specifications. The bandwidth is often expressed as a fraction of the centre frequency, which is midway between the extremities of the bandwidth.

1.2.6 Noise temperature

The minimum signal power that a receiving circuit can detect is limited by noise, which is passed to the circuit from the antenna, and generated by the circuit itself. The noise power N, appearing at the antenna's terminals, is produced thermally within the antenna structure and comes from electromagnetic noise sources in the antenna's environment.

The *antenna noise temperature* T_{A}, is the temperature required to cause a resistor to generate a thermal noise equal to N [Collin and Zucker, 1969a, Sec. 4.8]. Temperature T, is related to thermal noise power by *Nyquist's formula*

$$N = kT\,\Delta f \tag{1.25}$$

where k is Boltzmann's constant and Δf is the bandwidth of the system.

For an antenna surrounded by a distribution of distant, uncorrelated, electromagnetic noise sources, the total noise power received by the antenna is the sum of the noise powers received from each source. The noise power received from a source is proportional to the gain of the antenna in the direction of the source. By appealing to

(1.25), the noise distribution can be replaced by a *brightness temperature* distribution $T_b(\theta;\phi)$, such that, for a lossless antenna, the antenna temperature is given by [Rusch and Potter, 1970, Sec. 153]

$$T_A = \int_0^{2\pi}\int_0^{\pi} G(\theta;\phi)\,T_b(\theta;\phi)\sin\theta\,d\theta\,d\phi \tag{1.26}$$

In this formulation, $T_b(\theta;\phi)$ depends only on the noise sources and is independent of the directional characteristics of the antenna. If the antenna is not lossless, the thermal noise associated with R_{loss} also contributes to the antenna noise temperature [Jordan and Balmain, 1968, p. 416].

The brightness temperature of an object is equal to its ambient temperature T_0 only if it is a black body, implying that it absorbs all radiation incident upon it. If it absorbs a fraction α of the radiation power incident upon it, its brightness temperature is αT_0 [Collin and Zucker, 1969a, p. 119]. Non-thermal noise sources usually have a brightness temperature greater than T_0.

1.3 TYPES OF ANTENNA

There are as many different antennas as there are ways to radiate or receive an electromagnetic wave. Different antennas are suited for operation at different frequencies, have different directional characteristics and have different impedances. Antennas can be loosely categorized into four groups with different modes of operation: current elements, travelling wave, arrays and apertures [Rudge *et al.*, 1982, p. 2]. In this section the salient characteristics and method of analysis of each of these groups is briefly described.

1.3.1 Current elements

Current element antennas are less than a wavelength in size and are employed for radio frequencies up to about 1 GHz (that is, for λ greater than about 0.3 m).

The simplest radiating source is the *elemental dipole* (also called the Hertzian dipole or electric current element), which can be physically represented by a short, thin wire carrying a uniform current distribution. From Maxwell's equations, a current I in an elemental dipole of length dl, centred on the origin of a spherical coordinate system (Fig. 1.2) and parallel to the $\hat{\mathbf{z}}$ direction, produces a far field pattern of [Collin and Zucker, 1969a, Sec. 2.2]

$$\mathbf{E}(\theta;\phi) = j\frac{60\pi I\,dl}{\lambda r}e^{-j2\pi r/\lambda}\sin\theta\,\hat{\boldsymbol{\theta}} \tag{1.27}$$

A *short element antenna* is one whose length does not exceed about $\lambda/10$ and whose current distribution is approximately uniform. To calculate the radiation pattern of a current carrying wire of any length, one can consider it to be made up of elemental dipoles, and superimpose the contributions of each. For a thin wire with a current distribution of $I(z)$, the far field pattern is [Rudge *et al.*, 1982, p. 52]

$$\mathbf{E}(\theta;\phi) = j\frac{60\pi\sin\theta}{\lambda r}\int I(z)e^{-j2\pi r/\lambda}\,dz\;\hat{\boldsymbol{\theta}} \tag{1.28}$$

The most common type of current element antenna is a half wave dipole, which is a thin wire whose length at mid-band is $\lambda/2$ and whose current distribution is approximately $I(z) = I_0\cos 2\pi z/\lambda$ [Rudge *et al.*, 1982, p. 52]. An elemental magnetic

dipole, or a loop antenna element, can be constructed from four elemental dipoles connected in series, forming a square. A finite sized loop can be considered a distribution (usually circular) of contiguous elemental dipoles. Other derivatives of the elemental dipole include cylindrical rod antennas, vertical radiators and monopoles [Jasik, 1961, Chap. 3].

1.3.2 Travelling wave antennas

The distribution of current on a current element antenna can be treated as the sum of two current waves travelling in opposite directions. A *travelling wave antenna* is one in which the currents can be represented by one or more waves, usually travelling in the same direction [Walter, 1965, p. 13]. They are typically between 1 and 10 wavelengths long, and operate at frequencies between 1 MHz and 10 GHz. There are two stages to the analysis of these antennas [Rudge *et al.*, 1982, Sec. 1.15]. Firstly, the manner in which the current wave propagates along the length of the antenna must be deduced, and secondly, the contribution of these currents to the radiated electromagnetic wave must be calculated.

One form of travelling wave antenna is a long wire, terminated with a matched impedance at one end, and driven at the other. Such an antenna has an approximately uniform current amplitude distribution, with a progressive phase lag [Walter, 1965, Sec. 8.2]. The far field pattern can be calculated using (1.28).

Several terminated long wire antennas can be judiciously oriented and connected, in series and/or in parallel, to reinforce the beams in one particular direction. The rhombic antenna [ITT, 1968, p. 25-9] is based on this principle. In other travelling wave antennas, such as dielectric rod or helical antennas, the wave travels at a speed slower than the speed of electromagnetic radiation in free space [Walter, 1965, Sec. 8.3].

1.3.3 Aperture antennas

In an *aperture antenna* the radiated electromagnetic fields can be considered to emanate from a physical opening called an *aperture* [Collin and Zucker, 1969a, Chap. 3], which can in practice be anywhere up to several thousand wavelengths across. To achieve a manageable size, aperture antennas are usually operated in microwave bands (which have wavelengths of less than about 0.3 m).

Analysis of an aperture antenna tends to differ from the analysis of other types of antenna, because the radiating field is usually considered to be produced by field elements, rather than by elemental dipoles. The *aperture plane* is a plane through which most of the radiation passes and is near to, or coincident with, the aperture [IEEE, 1984]. The radiation pattern is computed as the sum of the individual radiations from all field elements in the aperture plane. This is discussed in further detail in Chapter 2.

A horn antenna is used to increase the peak gain of the waveguide which feeds it [Collin and Zucker, 1969a, Sec. 15.1]. A reflector antenna can be designed to produce a radiation pattern of almost any shape [Wood, 1980, p. 4], and is often highly directional. A dielectric, or metal plate, lens antenna also lends itself to aperture analysis [Silver, 1949, Chap. 11].

1.3.4 Arrays

An *array of antennas* consists of several individual antennas, called elements, positioned in a geometrical arrangement which is either regular or irregular. Almost any kind of antenna can be used as an array element, but all the element antennas in single array are usually identical, or at least similar [Ma, 1974, p. 1].

The radiation pattern of the array is the sum of contributions from each of the element antennas [Blake, 1984, Sec. 5.1]. The radiating field contribution of an element antenna depends on its position, its radiation pattern, the amplitude and phase of the signal feeding it and the mutual coupling between it and the other element antennas.

A phased array is one in which the the direction of the main beam is scanned by electronically altering the relative phase of the signal fed to each element antenna. In adaptive arrays, both the relative amplitude and the phase of the signal for each element antenna are controlled (often by computer), to produce a desired time varying radiation pattern [Rudge *et al.*, 1982, Sec. 1.13].

1.4 RADIO WAVE PROPAGATION OVER THE EARTH

Analysis of the propagation of electromagnetic waves between two antennas is straightforward when they are are within, and far (in comparison to the distance between the antennas) from the edges of, a volume of free space. The transmitting antenna radiates an electromagnetic wave which can be detected by the receiving antenna. Rays from one antenna to the other are straight because the refractive index is constant throughout the volume. The power per unit area of the wave at the receiving antenna depends on the transmitted power and gain pattern of the transmitting antenna and on the distance between the two antennas.

On a dry, still day, when two antennas are close to each other, compared to the distance to the earth's surface and other obstacles, the earth's atmosphere behaves approximately like free space [David and Voge, 1969, Sec. 4.2]. However, the analysis of radio wave propagation over or through the earth is usually complicated by the nature of the transmission media involved [David and Voge, 1969; Picquenard, 1974]. Rays between the antennas can be reflected or refracted due to inhomogeneities in the atmosphere. The power at the receiving antenna can be reduced because of absorption or path obstruction. These effects vary irregularly with time and depend on the geographical location of the antennas. Some of the atmosphere's effects on radio wave propagation are illustrated in Figure 1.5 and are discussed in the following sections.

1.4.1 The terrain

The earth and sea have conductivities and dielectric constants which are considerably different from those of the air immediately above, and which vary from place to place over the earth's surface [David and Voge, 1969, Sec. 2.2].

When the transmitting and receiving antennas are both near the earth's surface, the part of a radio wave which propagates parallel, and close (at a height of less than a wavelength), to the earth's surface is called a *surface wave*. Acting in a similar way to a transmission line, the ground guides the wave and absorbs some of its energy [Jordan and Balmain, 1968, p. 629]. Over long distances, when the antennas are not in line of sight, further attenuation is caused by the effect of the curvature of the earth. This attenuation is less for low frequency waves because they diffract around the earth more strongly than higher frequency waves [David and Voge, 1969, Sec. 4.4.2].

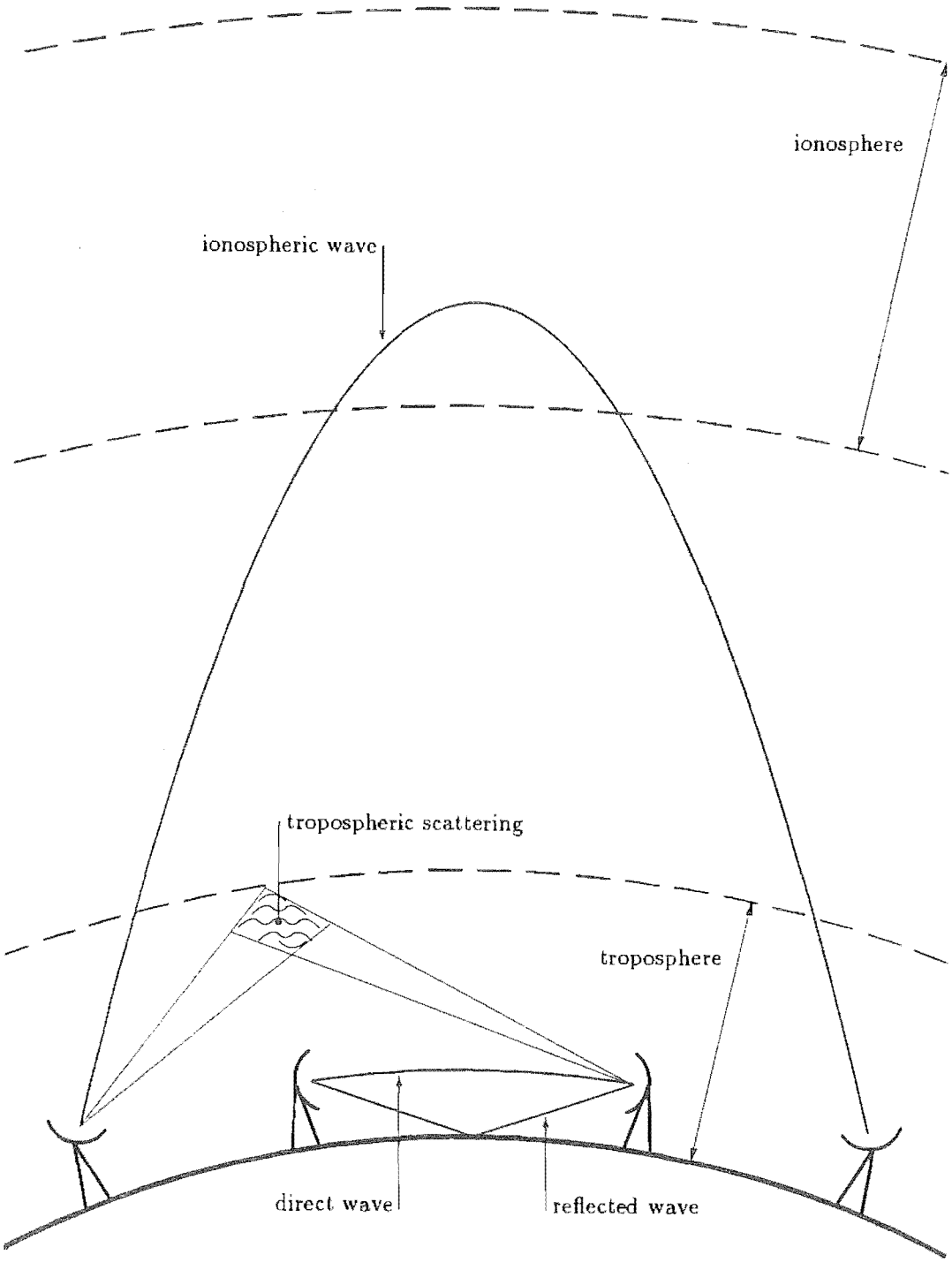


Figure 1.5 Some of the possible propagation paths of energy between two antenna sites.

The earth's surface can also act as a reflector. The part of an electromagnetic wave which is reflected by the ground is called a *ground reflected wave* and suffers attenuation and a phase shift dependent on the local properties of the terrain [Picquenard, 1974, Sec. 4.1].

1.4.2 The troposphere

The *troposphere* is the part of the earth's atmosphere in which the average temperature decreases with altitude, clouds form and convection is active [IEEE, 1984]. It is the layer up to about 10 km above the earth's surface. Its refractive index, which is always close to unity, varies randomly with position and time, about a mean which is related to altitude. A *standard atmosphere* has a refractive index gradient, with altitude, of -0.039×10^{-6} per metre [Picquenard, 1974, Sec. 2.1.1].

The parts of a radio wave which are radiated obliquely into a standard atmosphere undergo progressive refraction and its rays curve downwards, causing it to follow the curvature of the earth to an extent. This means that the *direct wave* between a transmitting and a receiving antenna usually follows a curved path. The length and shape of the path changes due to the slow temporal variations in atmospheric conditions.

The curved rays, associated with a direct wave propagating through the troposphere, can be geometrically transformed to produce straight rays above a model earth with an *equivalent earth radius*, implying an equivalent wave propagating through free space. For a standard atmosphere, the equivalent earth radius is approximately equal to 4/3 of the radius of the earth [Picquenard, 1974, Sec. 2.1.3.2].

Atmospheric conditions are occasionally such that the variation of refractive index with height causes a *duct* [Hall, 1979, Sec. 2.5], in which the rays are refracted around the earth, thus extending the range of the direct wave to over the horizon. This can only be sustained in a still atmosphere.

Tropospheric scattering is a form of wave propagation attributed to scattering from random irregularities in the index of refraction. These irregularities are caused by turbulence in the atmosphere [Panter, 1972, p. 2]. Tropospheric scattering occurs for frequencies between 0.1 and 10 GHz and considerably attenuates the field. The intensity of the scattered field has a slow seasonal variation superimposed on fast fluctuations having periods of as low as 0.1 second [Panter, 1972, Sec. 12.3.1].

Water vapour and oxygen in the troposphere absorb microwave energy, with the attenuation peaking at about 22 and 200 GHz for water, and 60 and 120 GHz for oxygen [David and Voge, 1969, Sec. 5.4]. The attenuation changes with frequency and with the density of water or oxygen over the propagation path. Particles, especially water droplets, in the atmosphere cause radio wave scattering. For frequencies below 50 GHz, this can produce higher attenuation than that from absorption [Hall, 1979, Sec. 3.4]. The attenuation due to scattering increases with frequency and with rainfall rate.

1.4.3 The ionosphere

The ionosphere is in the upper atmosphere, between about 100 and 600 km. The sun's radiation (and to a lesser degree, falling meteorites) causes the very low pressure gases to ionize. This effect is linked with the solar intensity and therefore varies greatly between day and night and is significantly affected by solar storms [Picquenard, 1974, Sec. 6.1.2].

The layers of charged particles are conductive at frequencies lower than about 500 kHz, so waves of these frequencies are reflected by the ionosphere [Picquenard, 1974, Sec. 6.2.1]. Waves with a frequency of between 1.5 and 30 MHz penetrate into the ionized layers by several wavelengths and are progressively refracted in a manner which can be equivalent to a reflection. Waves which are reflected or refracted by the ionosphere are called *sky waves*, and can be propagated well over the horizon [David and Voge, 1969, Sec. 6.2.3].

Radio waves with frequencies higher than 100 MHz can pass through the ionosphere. The presence of the earth's magnetic field in the ionized region causes the direction of polarization, of a traversing linearly polarized wave, to be rotated: an effect called *Faraday rotation* [Picquenard, 1974, Sec. 5.4]. The ionosphere can also absorb an appreciable fraction of the wave's power.

1.4.4 Interference

When two or more waves (for example, a direct wave and a ground reflected wave) arrive at a receiving antenna, they form an interference pattern because of their differing path lengths and phase delays [ITT, 1968, p. 21-10]. Therefore, the received power per unit area depends on the receiving antenna position and on the atmospheric conditions at the time. The interference effects can be reduced if the radiation pattern of either the transmitting antenna or the receiving antenna is directional enough to discriminate between the different waves.

There are several sources of electromagnetic noise which can interfere, at the receiving antenna, with the transmitted wave. *Atmospherics* originate from storms or electrical discharges between clouds, and have a brightness temperature ranging from about 10^{15} K at 100 kHz to about 100 K at 2 MHz [Jordan and Balmain, 1968, p. 413]. *Artificial noise* is produced by electrical equipment, and can have a relatively high level near towns and cities. Artificial noise power levels decrease with increasing frequency [ITT, 1968, p. 27-4]. *Extraterrestrial noise* comes from the sun and other stars, and is due to thermal radiation. Cosmic noise varies from about 10^5 K at 30 MHz to about 1 K at 1 GHz [Jordan and Balmain, 1968, p. 415]. Particles in the atmosphere which absorb radio wave energy are also a source of thermal noise radiation (cf. Sec. 1.2.6). At the peak absorption frequencies of water vapour and oxygen (see the last paragraph in Sec. 1.4.2), the brightness temperature can be between 20 and 290 K [Hall, 1979, p. 77; Jordan and Balmain, 1968, p. 415].

1.5 SUMMARY

Radio waves are a subset of electromagnetic waves, which consist of time varying electric and magnetic fields. Maxwell's equations (1.9) relate these fields to each other, to electromagnetic sources (time varying currents and charges) and to the constitutive parameters of the media in which the waves exist. The propagating wave nature of radio waves can be demonstrated by rearranging Maxwell's equations to form the wave equations (1.14). A radio wave can be completely described by the amplitude and phase of each vector component of the electric field at each point in space.

The purpose of an antenna is to transmit or to receive radio waves. In the far field region the angular distribution of a transmitted wave does not depend upon distance from the antenna. The far field directional characteristics of an antenna, while it is transmitting, are the same as for when the antenna is receiving. Important electrical

parameters of an antenna at a given frequency are its radiation pattern, impedance, radiation efficiency and noise temperature (which depends on the environment in which the antenna is operating).

Propagation of radio waves in the vicinity of the earth is affected by the atmosphere and the terrain. The waves can be reflected from the surface of the earth and from the ionosphere. They can be refracted by the troposphere and the ionosphere, and diffracted around the surface of the earth. Particles in the air can absorb and scatter energy from a traversing radio wave. Most of these effects vary, in an irregular manner, with time and with geographical location.

Perhaps the main application of antennas is for communication systems. An oscillating current used to drive a transmitting antenna radiates an electromagnetic wave which in turn induces, on a receiving antenna, an oscillating current of the same frequency, and proportional amplitude and phase. It follows that any variations in the transmitting current produces proportional variations in the receiving current. By encoding information into these variations, the information can be communicated as the modulation of the radio wave.

Other applications of antennas include: radar, in which information about a target is abstracted from a radio wave reflected off the target; radio astronomy, which is concerned with the measurement of natural radio signals from cosmic sources; airborne and marine navigational aid systems; and meteorological aids.

CHAPTER 2

HIGH GAIN REFLECTOR ANTENNAS

A reflector antenna consists of a feed and one or more reflectors. The feed is a small antenna (or an array of small antennas) which, on transmission, acts as the source of an electromagnetic wave. The reflectors are fabricated from highly conductive materials, so that electromagnetic waves incident upon them are reflected with minimum energy loss. The feed is designed (or chosen) to direct most of its electromagnetic power towards the reflectors. The reflectors are designed (i.e. shaped) to redirect the electromagnetic power in predominantly one direction, thus producing a highly directive radiation pattern.

Section 2.1 outlines approximate methods for analysing reflector antennas, and so provides an understanding of the way in which reflectors affect electromagnetic waves. The performance of a high gain antenna is strongly influenced by its gain pattern, which is discussed in Section 2.2. Different configurations of reflector antennas and their relative merits are discussed in Section 2.3, and in Section 2.4 some applications for high gain antennas are mentioned.

2.1 ANALYSIS OF SCATTERING FROM REFLECTORS

Assume that the *source field* $\mathbf{E}_0$, from a feed which is far from any other body, is known, either from direct measurement, or from theoretical analysis. When a conducting body is placed in the vicinity of the feed, surface currents $\mathbf{J}_s$ are induced on the body and radiate a *scattered field* $\mathbf{E}_s$. The total field is equal to $[\mathbf{E}_0 + \mathbf{E}_s]$ [James, 1986, Sec. 2.4.1]. The problem of analysing scattering from reflectors is to calculate the total field, given $\mathbf{E}_0$ and the shape of the conducting body.

Exact descriptions of scattering from reflectors can in principle be deduced from Maxwell's equations (Sec. 1.1.2). However, because these tend to be unmanageable in practice [James, 1986, Sec. 2.4.1], one is usually forced to resort to approximate approaches. The following sections summarize ray optics (Sec. 2.1.1), current-integration (Sec. 2.1.2) and field integration (Sec. 2.1.3), which are all methods of analysing reflector antennas in ways that are inexact but are nevertheless very useful.

2.1.1 Ray optical methods

Classical geometrical optics is that branch of optics which corresponds to the limiting case of $k \to \infty$ [Born and Wolf, 1970, Sec. 3.1], in which the energy of a light wave travels through an isotropic medium along straight rays, and diffraction effects are negligible. However, classical geometrical optics neglects wavelength, phase and the vector nature of electromagnetic waves [Collin and Zucker, 1969b, Sec. 16.1], which are all important for antenna analysis at radio frequencies. The following sections

demonstrate how classical geometrical optics is extended to include these factors. The theory of uniform plane waves (Sec. 2.1.1.1), and their reflection from plane surfaces (Sec. 2.1.1.2), leads to the geometrical optics method (Sec. 2.1.1.3). The result is a ray tracing method which is straightforward to apply to reflector analysis (Sec. 2.1.1.4). The main deficiency of the geometrical optics method is that it neglects diffraction effects. These effects are accommodated by the geometrical theory of diffraction, which is outlined in Section 2.1.1.5.

2.1.1.1 Uniform plane waves

The simplest solution to Maxwell's equations (1.9) is a *uniform plane wave*. Such an electromagnetic wave has an electric field which is constant over a plane. To satisfy (1.9), in free space, a uniform plane wave must be of the form [Ramo *et al.*, 1965, Sec. 6.02; Collin and Zucker, 1969a, Sec. 1.5]

$$\begin{aligned} \mathbf{E}(\mathbf{r}) &= \mathbf{e}_0 e^{-jk\mathbf{r}\cdot\hat{\mathbf{s}}_0} \\ \mathbf{e}_0 \cdot \hat{\mathbf{s}}_0 &= 0 \end{aligned} \tag{2.1}$$

where $\mathbf{r}$ is the position vector of an arbitrary point in space, $\hat{\mathbf{s}}_0$ is the unit vector normal to the plane and $\mathbf{e}_0$ is the (constant complex vector) value of the electric field at $\mathbf{r} = 0$. Substituting (2.1) into (1.13) gives

$$\mathbf{H}(\mathbf{r}) = \left(\frac{\epsilon}{\mu}\right)^{1/2} \hat{\mathbf{s}}_0 \times \mathbf{E}(\mathbf{r}) \tag{2.2}$$

Therefore, at all points in space, $\mathbf{H}$ is always proportional to $\mathbf{E}$. Furthermore, $\mathbf{E}$ and $\mathbf{H}$ are perpendicular to each other and to $\hat{\mathbf{s}}_0$.

An expression for the complex Poynting vector at each point in a uniform plane wave is obtained by substituting (2.2) into (1.16):

$$\mathbf{P}(\mathbf{r}) = \frac{1}{2}\left(\frac{\epsilon}{\mu}\right)^{1/2} (\mathbf{E}(\mathbf{r}) \cdot \mathbf{E}^*(\mathbf{r}))\, \hat{\mathbf{s}}_0 \tag{2.3}$$

This shows that the average power flow at each point in space is proportional to the square of the electric field magnitude (see (1.6)) and is in the direction of $\hat{\mathbf{s}}_0$. Since $\hat{\mathbf{s}}_0$ is a constant, rays (which are parallel to $\mathbf{P}$ (Sec. 1.1.5)) representing a uniform plane wave are always straight lines.

2.1.1.2 Uniform plane wave reflection

Figure 2.1 shows a ray of a plane wave incident upon the infinitely large plane surface of a perfect conductor. Since the reflected field is also a plane wave, the total field is given by

$$\mathbf{E}(\mathbf{r}) = \mathbf{E}_i(\mathbf{r}) + \mathbf{E}_r(\mathbf{r}) \tag{2.4}$$

where the subscripts 'i' and 'r' denote, respectively, the incident and reflected fields. From the definition of a plane wave (2.1), the vector description of a plane surface and the boundary condition for a conductor (1.12), the following conditions must be satisfied on the conductor surface S:

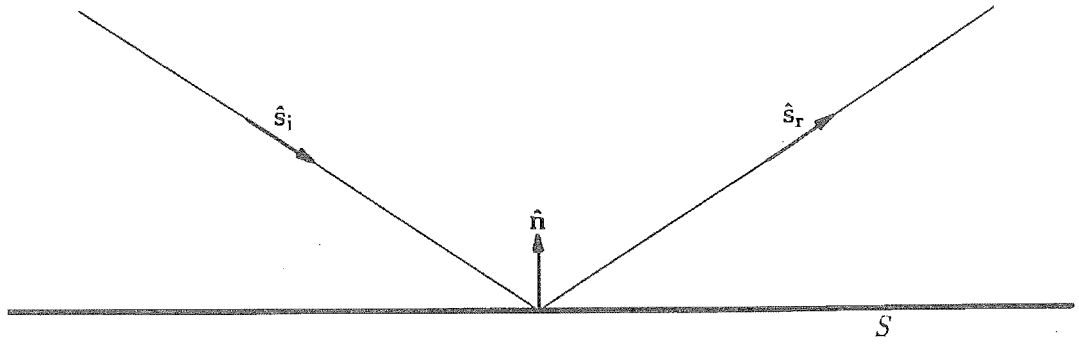


Figure 2.1 Ray diagram showing the reflection of a plane wave from a plane reflecting surface.

$$\begin{aligned} \mathbf{E}_i(\mathbf{r}') &= \mathbf{e}_i e^{-jk\hat{\mathbf{s}}_i\cdot\mathbf{r}'} \\ \mathbf{e}_i\cdot\hat{\mathbf{s}}_i &= 0 \\ \mathbf{E}_r(\mathbf{r}') &= \mathbf{e}_r e^{-jk\hat{\mathbf{s}}_r\cdot\mathbf{r}'} \\ \mathbf{e}_r\cdot\hat{\mathbf{s}}_r &= 0 \\ \hat{\mathbf{n}}\cdot\mathbf{r}' &= 0 \\ \hat{\mathbf{n}}\times(\mathbf{E}_i(\mathbf{r}')+\mathbf{E}_r(\mathbf{r}')) &= 0 \end{aligned} \tag{2.5}$$

where $\hat{\mathbf{n}}$ is the unit normal to the reflector surface, pointing away from the conductor, and $\mathbf{e}_i$, $\hat{\mathbf{s}}_i$, $\mathbf{e}_r$ and $\hat{\mathbf{s}}_r$ are constants defining the plane waves. An arbitrary point on the reflector surface is denoted by the position vector $\mathbf{r}'$. Solving these equations, and expressing the reflected field in terms of the incident field, yields [Silver, 1949, Sec. 5.3]

$$\begin{aligned} \hat{\mathbf{s}}_r &= \hat{\mathbf{s}}_i - 2(\hat{\mathbf{n}}\cdot\hat{\mathbf{s}}_i)\,\hat{\mathbf{n}} \\ \mathbf{e}_r &= -\mathbf{e}_i + 2(\hat{\mathbf{n}}\cdot\mathbf{e}_i)\,\hat{\mathbf{n}} \end{aligned} \tag{2.6}$$

assuming that $\hat{\mathbf{s}}_i$ is not parallel to the reflecting surface.

It follows from the first equation of (2.6) that

$$\begin{aligned} \hat{\mathbf{n}}\times\hat{\mathbf{s}}_r &= \hat{\mathbf{n}}\times\hat{\mathbf{s}}_i \\ \hat{\mathbf{n}}\cdot\hat{\mathbf{s}}_r &= -\hat{\mathbf{n}}\cdot\hat{\mathbf{s}}_i \end{aligned} \tag{2.7}$$

which gives the two *laws of reflection*:

1. $\hat{\mathbf{s}}_i$, $\hat{\mathbf{s}}_r$ and $\hat{\mathbf{n}}$ all lie in a single plane, called the *plane of incidence*.
2. The angle between $\hat{\mathbf{n}}$ and $-\hat{\mathbf{s}}_i$ equals the angle between $\hat{\mathbf{n}}$ and $\hat{\mathbf{s}}_r$.

The surface current (Sec. 1.1.4), induced by the electromagnetic wave on the reflecting surface, is given by the substitution of (2.2) and (2.4) into the third equation of (1.12):

$$\mathbf{J}_s(\mathbf{r}') = \left(\frac{\epsilon}{\mu}\right)^{1/2}\hat{\mathbf{n}}\times(\hat{\mathbf{s}}_i\times\mathbf{E}_i(\mathbf{r}')+\hat{\mathbf{s}}_r\times\mathbf{E}_r(\mathbf{r}')) \tag{2.8}$$

After substituting the first and third equation of (2.5), and (2.6), into the above equation, it simplifies to [Silver, 1949, p. 134]

$$\mathbf{J}_s(\mathbf{r}') = 2\left(\frac{\epsilon}{\mu}\right)^{1/2} \hat{\mathbf{n}} \times (\hat{\mathbf{s}}_i \times \mathbf{E}_i(\mathbf{r}')) \tag{2.9}$$

This result is required for the physical optics method (Sec. 2.1.2.3).

2.1.1.3 Geometrical optics (GO)

By extension of classical geometrical optics theory to radio waves, the *geometrical optics (GO) approximation*, to an electromagnetic field, has the following form:

$$\mathbf{E}(\mathbf{r}) = \mathbf{e}(\mathbf{r})e^{-jks(\mathbf{r})} \tag{2.10}$$

where $\mathbf{e}(\mathbf{r})$ is a complex vector function of position and $s(\mathbf{r})$ is an independent, real scalar function. The GO approximation satisfies Maxwell's equations, in free space, if the following constraints are imposed [Rudge *et al.*, 1982, Sec. 2.2.2]:

$$\begin{aligned} \nabla s \cdot \nabla s &= 1 \\ \nabla s \cdot \mathbf{e} &= 0 \\ \nabla^2 s\, \mathbf{e} + 2(\nabla s \cdot \nabla)\mathbf{e} &= 0 \end{aligned} \tag{2.11}$$

with the further provisos:

1. $k \to \infty$ (i.e. the GO approximation is a high frequency approximation).

2. Changes in $\mathbf{r}$, over distances of the order of a wavelength, result in small relative changes in $\mathbf{e}(\mathbf{r})$ [Born and Wolf, 1970, p. 111 and p. 121].

The surfaces in space formed by constant $s(\mathbf{r})$ are called *geometrical wavefronts.*

The first equation of (2.11) ensures that $\nabla s(\mathbf{r})$ is everywhere a unit vector. Comparison of (2.10), and the second equation of (2.11), with (2.1) shows the similarity between the GO and plane wave fields. In a region of space local to $\mathbf{r}$, the GO field can be equated to a plane wave field defined by

$$\hat{\mathbf{s}}_0 = \nabla s(\mathbf{r}) \quad \text{and} \quad \mathbf{e}_0 = \mathbf{e}(\mathbf{r})e^{-jk(s(\mathbf{r})-\mathbf{r}\cdot\nabla s(\mathbf{r}))} \tag{2.12}$$

Substituting (2.12) into (2.3) reveals that the Poynting vector is everywhere parallel to $\nabla s(\mathbf{r})$. Therefore GO rays are normal to the wavefronts. In free space, these rays are straight lines [Born and Wolf, 1970, Sec. 3.2.1].

For a known field incident upon a conductor, an approximation to the reflected field can be calculated using the *geometrical optics method* which assumes that, local to each point on the surface of the conductor, the incident field behaves as if it were part of a plane wave, and the conductor surface behaves as if it were part of an infinite plane [James, 1986, p. 98]. Therefore, the local interaction between an incident field and a conducting surface can be approximated by (2.6), which embodies the laws of reflection. Because it is constrained by (2.11), the reflected field, at a point in space not local to the conductor surface, is dependent upon the incident field and on the curvature of the conductor. This is explained in the following section.

2.1.1.4 Ray tracing

Ray tracing is a practical application of the geometrical optics method to the analysis of scattering from reflectors. A computational feature of the geometrical optics method is that $s(\mathbf{r})$ (or equivalently, rays) can be evaluated (or traced) independently of $\mathbf{e}(\mathbf{r})$. The field values at all points along the rays can be subsequently calculated.

The last equation of (2.11) ensures that the amplitude of the GO field is such that energy is conserved along pencils of rays [Born and Wolf, 1970, footnote on p. 118]. It also guarantees that the polarization of the field is constant along a ray in free space [Born and Wolf, 1970, p. 49]. Therefore, the phase of each vector component of $\mathbf{e}$, in (2.10), is constant along a ray, leaving the phase of each vector component of $\mathbf{E}$ to vary according to $ks(\mathbf{r})$. Since ∇s is a unit vector in the direction of the ray (see Sec. 2.1.1.3), the phase difference between two points $\mathbf{r}_1$ and $\mathbf{r}_2$ on a ray is equal to k times the length l along the ray between them.

It follows that, for a ray in free space, the field at $\mathbf{r}_2$ can be expressed in terms of the field at $\mathbf{r}_1$ [James, 1986, p. 102] by

$$\begin{aligned} \mathbf{E}(\mathbf{r}_2) &= \mathbf{E}(\mathbf{r}_1)\left(\frac{dA_1}{dA_2}\right)^{1/2} e^{-jkl} \\ \text{where } \mathbf{r}_2 &= \mathbf{r}_1 + l\nabla s(\mathbf{r}_1) \end{aligned} \tag{2.13}$$

and dA_1 and dA_2 are the cross-sectional areas, at positions $\mathbf{r}_1$ and $\mathbf{r}_2$ respectively, of a surrounding pencil of rays. The cross-sectional areas are taken over wavefront surfaces.

When an incident ray, of direction ∇s_i, intersects a reflecting surface at point $\mathbf{r}'$, with a field value of $\mathbf{E}_\mathrm{i}(\mathbf{r}')$, the ray is reflected in the direction ∇s_r and has a field value $\mathbf{E}_\mathrm{r}(\mathbf{r}')$. The fields are related by (cf. (2.6))

$$\begin{aligned} \nabla s_\mathrm{r} &= \nabla s_\mathrm{i} - 2(\hat{\mathbf{n}} \cdot \nabla s_\mathrm{i})\,\hat{\mathbf{n}} \\ \mathbf{E}_\mathrm{r}(\mathbf{r}') &= -\mathbf{E}_\mathrm{i}(\mathbf{r}') + 2(\hat{\mathbf{n}} \cdot \mathbf{E}_\mathrm{i}(\mathbf{r}'))\,\hat{\mathbf{n}} \end{aligned} \tag{2.14}$$

where $\hat{\mathbf{n}}$ is the unit normal to the reflecting surface at the point $\mathbf{r}'$.

Figure 2.2 illustrates the tracing of several rays. It is assumed that the field is known over a surface which completely encloses the feed. The direction of a ray, passing through a point A on this surface, is given by the Poynting vector of the field at A. Each ray comprising a pencil of rays originates from the circumference of a small area of the surface surrounding A. In free space, each of the rays is straight, and should the pencil intersect a conductor surface, each ray is independently reflected according to the laws of reflection (see Sec. 2.1.1.2). A given pencil may be reflected one or more times by reflectors, or may not be reflected at all. The value of the field at any point along the central ray is completely specified by (2.13) and (2.14). The value of the total field at any point in space is the sum of the fields of each ray passing through the point.

The ratio $dA_1 \,/\, dA_2$, which appears in (2.13), is often expressed in terms of l and the principal radii of curvature of the wavefront at $\mathbf{r}_1$, while the principal radii of a reflected wavefront are expressed in terms of the principal radii of both the incident wavefront and the reflecting surface, all taken at point $\mathbf{r}'$ [Rudge *et al.*, 1982, Sec. 2.2; James, 1986, Sec. 4.2]. This approach removes the need to trace all the individual rays comprising the pencil of rays.

Two deficiencies of the GO method are apparent from Figure 2.2. Firstly, it does not predict diffraction effects. This means that the GO field is zero in the *shadow region*,

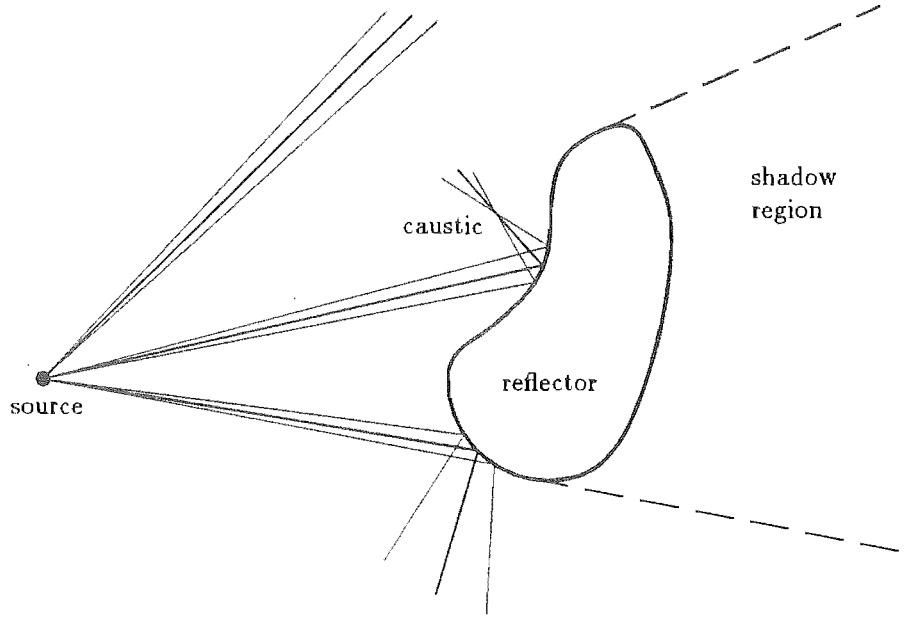


Figure 2.2 Examples of the geometry of ray tracing, from a known source distribution, to the vicinity of a conducting surface in free space.

which is on the far side of the reflector from the feed (Fig. 2.2). Secondly, a *caustic* is predicted at all points where pencils of rays have zero cross-sections. At these points the field predicted by the GO method is infinite, so alternative representations of the field must be employed. These deficiencies arise because, at the boundary of a shadow region and at caustics, the field changes rapidly with position, thereby invalidating assumptions underlying the derivation of (2.11) [Born and Wolf, 1970, p. 121].

2.1.1.5 Geometrical theory of diffraction (GTD)

The *geometrical theory of diffraction (GTD)* extends the GO approximation by postulating rays which account for diffraction as well as reflection. In the analysis of reflector antennas, the main kind of diffraction encountered is that at edges of conducting materials. The total field $\mathbf{E}(\mathbf{r})$ is given by

$$\mathbf{E}(\mathbf{r}) = \mathbf{E}_{\mathrm{g}}(\mathbf{r}) + \mathbf{E}_{\mathrm{d}}(\mathbf{r}) \tag{2.15}$$

where $\mathbf{E}_{\mathrm{g}}(\mathbf{r})$ is the GO field. The term $\mathbf{E}_{\mathrm{d}}(\mathbf{r})$ is proportional to $k^{-1/2}$ and is a high frequency approximation to the diffracted field. In free space, $\mathbf{E}_{\mathrm{d}}(\mathbf{r})$ behaves like the GO field, so its value at two points on a straight ray are related by (2.13).

For a known field incident upon a conductor, the edge diffracted field can be calculated using the *GTD method.* This method assumes that, local to each point on the edge of the conductor, the incident field behaves as if it were part of a plane wave, and the conductor edge behaves as if it were part of an infinite straight edge [James, 1986, p. 132].

A ray parallel to $\hat{\mathbf{s}}_{\mathrm{i}}$, incident upon an edge of a conductor gives rise to a cone of

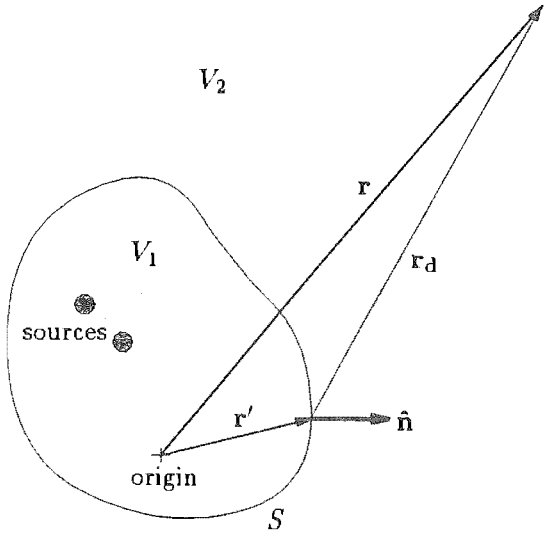


Figure 2.3 Geometry of integration methods for determining the field at point r.

diffracted rays, such that for each diffracted ray [James, 1986, Sec. 3.2]

$$\hat{s}_d \cdot \hat{d} = \hat{s}_i \cdot \hat{d} \tag{2.16}$$

where $\hat{s}_d$ is parallel to the diffracted ray and $\hat{d}$ is parallel to the conductor edge. The value of the field on the diffracted ray is proportional to the incident field and is a function of $\hat{s}_d$. The principles of ray tracing apply to GTD as well as to GO.

In contrast to GO, the GTD field does predict a field in shadow regions, but it still predicts an infinite field at caustics, and still gives invalid results near the boundaries of GO shadow regions [Wood, 1980, Sec. 2.3.2]. Several methods of correcting for these non-uniformities are reviewed by Arnold [1986].

2.1.2 Current-integration methods

Integration methods for calculating the field in a volume usually require more computational effort than ray tracing techniques, but they involve fewer approximations and therefore provide more accurate solutions. Unlike the GO method, integration methods account for diffraction effects.

The field scattered by a reflector can be predicted by current-integration methods when the currents on the reflector are known (Sec. 2.1.2.1). When evaluating the far field, certain approximations to the exact formulation are made (Sec. 2.1.2.2). There remains the problem of calculating the currents on the reflector surface. The most important components of these currents can be calculated acceptably accurately with the aid of the physical optics method (Sec. 2.1.2.3), supplemented by GTD (Sec. 2.1.1.5).

2.1.2.1 Surface current integration

Figure 2.3 depicts a surface S enclosing a volume V_1, containing all the electromagnetic sources. The volume V_2 encompasses all free space outside V_1. From Maxwell's equations, the electric field at any point r in V_2 can be expressed in terms of the magnetic

and electric fields over S by [Rusch and Potter, 1970, Appendix]

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi}\int_S \left\{(\hat{\mathbf{n}}\times\mathbf{E})\times\nabla\psi - j\omega\mu(\hat{\mathbf{n}}\times\mathbf{H})\psi + \frac{1}{j\omega\epsilon}[(\hat{\mathbf{n}}\times\mathbf{H})\cdot\nabla]\nabla\psi\right\} dS \quad (2.17)$$

where dS is an elemental area on S and

$$\psi(\mathbf{r}',\mathbf{r}) = \frac{e^{-jkr_d}}{r_d} \quad (2.18)$$

where $r_d = |\mathbf{r}_d|$ and $\mathbf{r}_d = (\mathbf{r}-\mathbf{r}')$. The operator ∇ operates on the coordinates of $\mathbf{r}'$, which is the position vector of an arbitrary point on S.

The field throughout V_2 is unaffected if the field and sources within V_1 are replaced by a null field within V_1 and equivalent sources on S [Collin and Zucker, 1969a, Sec. 3.3]. The surface electric and magnetic currents, $\mathbf{J}_s$ and $\mathbf{J}_{ms}$ respectively, comprising the equivalent sources must satisfy

$$\begin{aligned}\mathbf{J}_s(\mathbf{r}') &= \hat{\mathbf{n}}\times\mathbf{H}(\mathbf{r}')\\ \mathbf{J}_{ms}(\mathbf{r}') &= -\hat{\mathbf{n}}\times\mathbf{E}(\mathbf{r}')\end{aligned} \quad (2.19)$$

which follow from the boundary conditions (1.11). Substitution of (2.19) into (2.17) yields an expression for the field in V_2 in terms of equivalent current distributions over S:

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi}\int_S \left\{-\mathbf{J}_{ms}\times\nabla\psi - j\omega\mu\mathbf{J}_s\psi + \frac{1}{j\omega\epsilon}[\mathbf{J}_s\cdot\nabla]\nabla\psi\right\} dS \quad (2.20)$$

2.1.2.2 Approximations in the far field region

At distances far from the source, the expression (2.20) for the field can be simplified. When (2.18) is substituted into (2.20), the integrand on its right hand side becomes a power series in r_d^{-1} [Silver, 1949, p. 87]. Provided that $r_d \gg \lambda$, the terms of second and higher degree are negligible. Equation (2.20) then reduces to [Silver, 1949, p. 88]

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi}\int_S \left\{-jk\mathbf{J}_{ms}\times\hat{\mathbf{r}}_d - j\omega\mu\mathbf{J}_s - \frac{k^2}{j\omega\epsilon}[\mathbf{J}_s\cdot\hat{\mathbf{r}}_d]\,\hat{\mathbf{r}}_d\right\}\frac{e^{-jkr_d}}{r_d} dS \quad (2.21)$$

where $\hat{\mathbf{r}}_d$ is the unit vector parallel to $\mathbf{r}_d$.

From the definition of the far field region provided by Section 1.2.1, the angular distribution of the far field does not depend on distance from the source. Therefore, if the origin is close to S (as in Fig. 2.3) and $\mathbf{r}$ is in the far field region, $\mathbf{r}$ is approximately parallel to $\mathbf{r}_d$. The effect of the difference in length between $\mathbf{r}$ and $\mathbf{r}_d$ is negligible for the amplitude of an arbitrary vector component of the far field, but can affect its phase significantly. This means that the following approximations can be made in the far field region [Silver, 1949, Sec. 3.11]:

$$\begin{aligned} r_d &= \begin{cases} r & \text{except in phase expressions}\\ r - \mathbf{r}'\cdot\hat{\mathbf{r}} & \text{in phase expressions}\end{cases}\\ \hat{\mathbf{r}}_d &= \hat{\mathbf{r}}\end{aligned} \quad (2.22)$$

where $\hat{\mathbf{r}}$ is a unit vector parallel to $\mathbf{r}$.

When the approximations in (2.22) are incorporated into (2.21) the expression for the far field becomes [Collin and Zucker, 1969a, Sec. 2.5]

$$\mathbf{E}(\mathbf{r}) = \frac{-jk}{4\pi r}e^{-jkr}\int_S \left\{\mathbf{J}_{ms}\times\hat{\mathbf{r}} + \left(\frac{\mu}{\epsilon}\right)^{1/2}[\mathbf{J}_s - (\mathbf{J}_s\cdot\hat{\mathbf{r}})\,\hat{\mathbf{r}}]\right\} e^{jk\mathbf{r}'\cdot\hat{\mathbf{r}}} dS \quad (2.23)$$

Local to $\mathbf{r}$ this represents a plane wave (Sec. 2.1.1.1) travelling away from the origin. Therefore, the radial component of the far field of a source is zero. The amplitudes of the tangential components of the far field decay with the inverse of distance from the source and vanish infinitely far from the source.

2.1.2.3 Physical optics

Equations (2.20) and (2.23) may be applied to the analysis of fields scattered from a conducting surface, because the field inside a perfect conductor is zero (Sec. 1.1.3). The surface S (Fig. 2.3) is taken to be the surface of the conductor, and the surface currents are those satisfying the boundary condition for a perfect conductor (1.12). Provided these surface currents are known exactly, the field reradiated by the reflector can be calculated exactly from (2.20). However, to know the surface currents, one must first know the total field at the surface.

The *physical optics (PO) method* assumes that each point on a conducting surface behaves locally as if it were part of an infinite plane. As illustrated in Figure 2.4, the reflector surface can be divided into two regions. S_1 is the region directly illuminated by the feed and S_2 is in the geometrical shadow of the feed. The *physical optics (PO) approximation* to the current induced on the surface of a conductor at point $\mathbf{r}'$ is (cf. (2.9))

$$\mathbf{J}_s(\mathbf{r}') = \begin{cases} 2\,\hat{\mathbf{n}} \times \mathbf{H}_i(\mathbf{r}') = 2\left(\dfrac{\epsilon}{\mu}\right)^{1/2} \hat{\mathbf{n}} \times (\hat{\mathbf{s}}_i \times \mathbf{E}_i(\mathbf{r}')) & \text{on } S_1 \\ 0 & \text{on } S_2 \end{cases} \tag{2.24}$$

where $\mathbf{E}_i$ is the field incident upon the conductor. Because this is a physical current, $\mathbf{J}_{ms} = 0$ and (2.20) can be utilized to calculate the reflected field. If only the far field is required, (2.23) can be used.

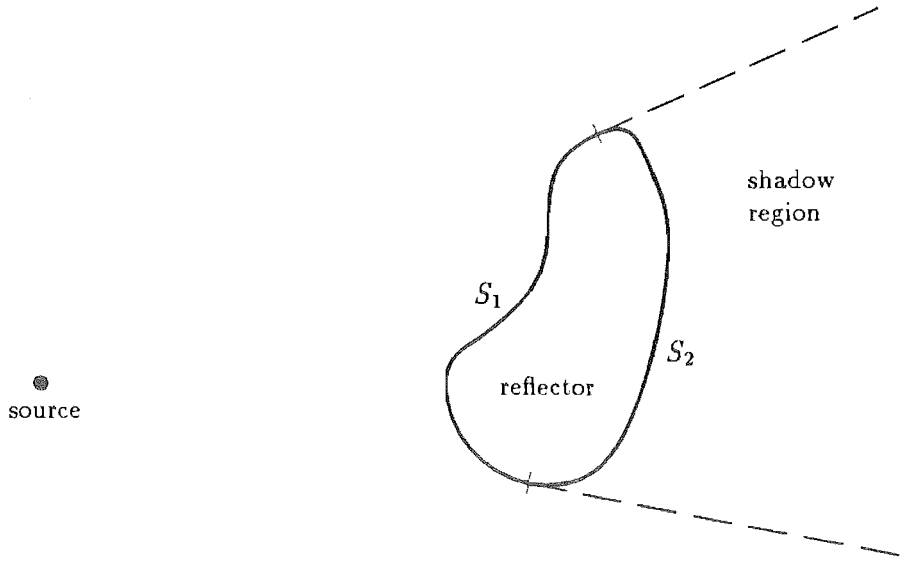


Figure 2.4 Geometry for the physical optics method.

On the illuminated side of the conductor, PO gives a more accurate result than GO, because PO allows for diffraction and does not predict caustics. However, it cannot be expected to provide accurate predictions of the field in or near the geometrical shadow region, because it neglects any currents in this region [James, 1986, Sec. 2.1.4].

2.1.3 Field integration methods

By definition, most of the radiation from a high gain antenna passes through its aperture plane (Sec. 1.3.3). Therefore, the analysis of these antennas can be simplified by invoking certain approximations. The electric far field radiated from the antenna can be expressed as an integral of the electric field on the aperture plane (Sec. 2.1.3.1). With an appropriate choice of coordinates, the far field pattern of the antenna is related to the field on the aperture plane by Fourier transformation (Sec. 2.1.3.2). The aperture field method (Sec. 2.1.3.3) provides a way of estimating the field on the aperture plane. Fourier transformation is invertible (Sec. 2.1.3.4) and can also be employed to calculate the field within the near field region (Sec. 2.1.3.5).

2.1.3.1 Equivalent currents on a plane

Figure 2.5(a) shows all of space divided into two half-space volumes V_1 and V_2, separated by an infinitely large plane surface S. The antenna is located within V_1, which is bounded by S, and a hemisphere of infinite radius. From (2.23), the field over the hemisphere is vanishingly small since the dimensions of the antenna are finite. The electromagnetic wave $(\mathbf{E}, \mathbf{H})$ is radiated by the antenna throughout both V_1 and V_2. It is required to evaluate the field in V_2.

Collin and Zucker [1969a, Sec. 3.4] show that the electromagnetic wave in V_2 is unaltered if the antenna in V_1 is removed and the following surface currents are placed over S (Fig. 2.5(b)):

$$\begin{aligned} \mathbf{J}_{ms}(\mathbf{r}') &= -2(\hat{\mathbf{n}} \times \mathbf{E}(\mathbf{r}')) \\ \mathbf{J}_s(\mathbf{r}') &= 0 \end{aligned} \tag{2.25}$$

Therefore, by appealing to (2.20), the electric field, at any point $\mathbf{r}$ in V_2, can be expressed exactly in terms of only the electric field over the plane surface S.

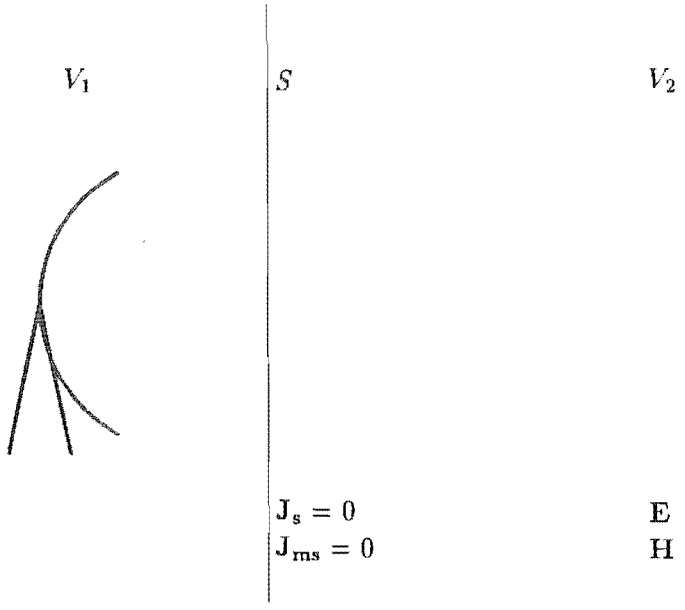
2.1.3.2 Fourier transformation

The fields radiated by the antenna (see Fig. 2.5(a)) and the equivalent surface currents (see Fig. 2.5(b)) are identical throughout V_2. Therefore, (2.23) can be invoked to predict the far field pattern of the antenna. However, the approximations (2.22) imply that the source of the field occupies a finite volume. Therefore, $\mathbf{J}_{ms}$ must be negligible outside a finite region S_0 of the infinite surface S. The far field of the source is thus given by substituting (2.25) into (2.23):

$$\mathbf{E}(\mathbf{r}) = \frac{jk}{2\pi r} e^{-jkr} \int_{S_0} \{(\hat{\mathbf{n}} \times \mathbf{E}(\mathbf{r}')) \times \hat{\mathbf{r}}\} \, e^{jk\mathbf{r}'\cdot\hat{\mathbf{r}}} dS \tag{2.26}$$

Figure 2.6 shows the plane surface S coinciding with the x, y plane, with S_0 centred on the origin. The field $\mathbf{E}(\mathbf{r}')$ on S is called the *aperture field*. A *radiation hemisphere* is defined to be that hemisphere, centred on the origin, existing for $z > 0$ and of radius R, where R is greater than the minimum far field distance. The angular distribution

(a)



(b)

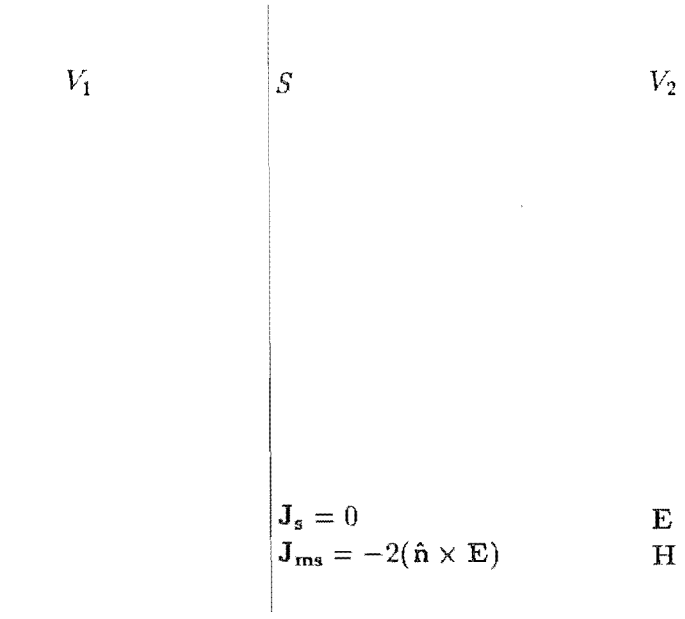


Figure 2.5 Equivalent currents on a plane surface: (a) antenna in volume V_1; (b) surface magnetic currents over the plane surface S. The antenna can be replaced by the surface magnetic currents without affecting the field in V_2.

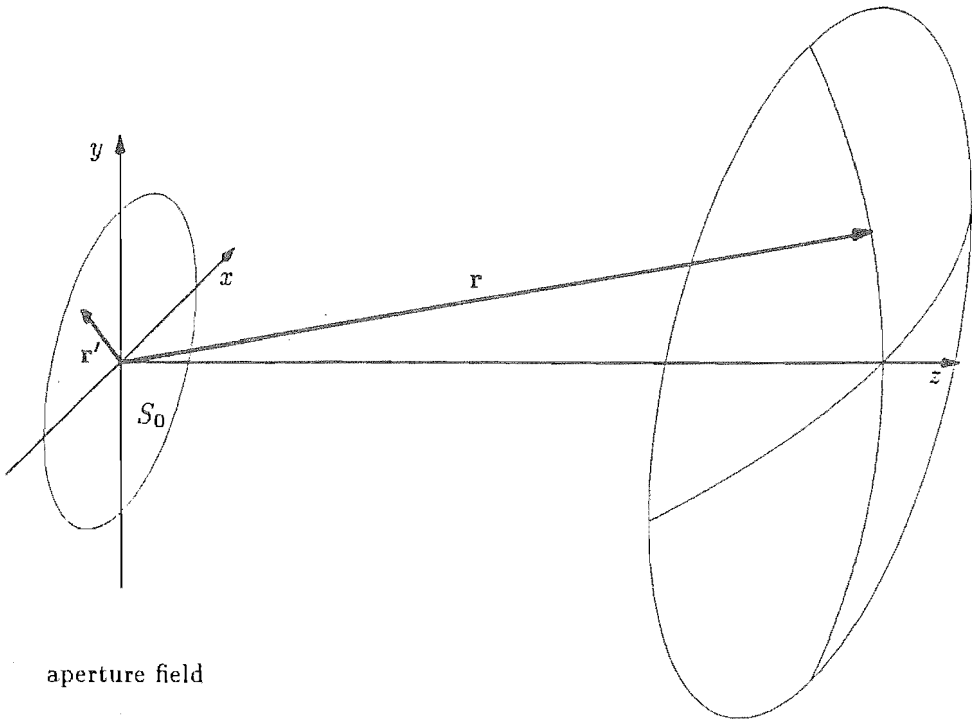


Figure 2.6 Geometry for establishing the Fourier transform relationship between the aperture and far fields.

of the radiated field $\mathbf{E}(\mathbf{r})$ on the surface of this radiation hemisphere is, by definition, the antenna's far field pattern (Sec. 1.2.2).

Position vectors and vector functions can be expressed in terms of their Cartesian components (see Sec. 1.1.1)

$$\begin{aligned} \mathbf{r}' &= x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}} \\ \mathbf{r} &= R(\sin\theta\cos\phi\,\hat{\mathbf{x}} + \sin\theta\sin\phi\,\hat{\mathbf{y}} + \cos\theta\,\hat{\mathbf{z}}) \\ \mathbf{E} &= E_x\,\hat{\mathbf{x}} + E_y\,\hat{\mathbf{y}} + E_z\,\hat{\mathbf{z}} \\ \hat{\mathbf{n}} &= \hat{\mathbf{z}} \end{aligned} \tag{2.27}$$

Only two parameters, the spherical ordinates θ and ϕ, are required to completely define $\mathbf{r}$, because $\mathbf{r}$ is confined to the two-dimensional surface of the radiation hemisphere. An alternative pair of parameters are u and v:

$$\begin{aligned} \mathbf{r} &= R\lambda\,(u\,\hat{\mathbf{x}} + v\,\hat{\mathbf{y}} + w(u,v)\,\hat{\mathbf{z}}) \\ \text{where}\quad u &= (\sin\theta\cos\phi)/\lambda = \hat{\mathbf{r}}\cdot\hat{\mathbf{x}}\,/\lambda \\ v &= (\sin\theta\sin\phi)/\lambda = \hat{\mathbf{r}}\cdot\hat{\mathbf{y}}\,/\lambda \\ w(u,v) &= (\cos\theta)/\lambda = \hat{\mathbf{r}}\cdot\hat{\mathbf{z}}\,/\lambda = ((1/\lambda^2) - u^2 - v^2)^{1/2} \end{aligned} \tag{2.28}$$

The positive square root in the definition of $w(u,v)$ is always taken, so that the z component of $\mathbf{r}$ is positive and therefore lies in V_2. The aperture field distribution and

the far field pattern can now be denoted by

$$\begin{aligned} \mathbf{E}(x,y) &= \mathbf{E}(x,y,0) \\ \text{and } \acute{\mathbf{E}}(u,v) &= \mathbf{E}(R\lambda u, R\lambda v, R\lambda w(u,v)) \end{aligned} \tag{2.29}$$

respectively.

The terms in (2.26) can be simplified with the aid of (2.27) and (2.28):

$$\begin{aligned} (\hat{\mathbf{n}} \times \mathbf{E}(\mathbf{r}')) \times \hat{\mathbf{r}} &= \lambda \left[w(u,v) E_x \hat{\mathbf{x}} + w(u,v) E_y \hat{\mathbf{y}} - (u E_x + v E_y) \hat{\mathbf{z}} \right] \\ jk\mathbf{r}' \cdot \hat{\mathbf{r}} &= j2\pi(xu + yv) \\ dS &= dx\, dy \end{aligned} \tag{2.30}$$

Substituting these terms into (2.26) and treating each Cartesian component of the field separately, gives

$$\begin{aligned} E_x(\mathbf{r}) = \acute{E}_x(u,v) &= jw(u,v)\frac{e^{-jkR}}{R} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E_x(x,y) e^{j2\pi(ux+vy)}\, dx\, dy \\ &= jw(u,v)\frac{e^{-jkR}}{R} \mathrm{FT}\{E_x(x,y)\} \\ E_y(\mathbf{r}) = \acute{E}_y(u,v) &= jw(u,v)\frac{e^{-jkR}}{R} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} E_y(x,y) e^{j2\pi(ux+vy)}\, dx\, dy \\ &= jw(u,v)\frac{e^{-jkR}}{R} \mathrm{FT}\{E_y(x,y)\} \\ E_z(\mathbf{r}) = \acute{E}_z(u,v) &= \frac{-u}{w(u,v)} \acute{E}_x(u,v) + \frac{-v}{w(u,v)} \acute{E}_y(u,v) \end{aligned} \tag{2.31}$$

where $\mathrm{FT}\{\cdot\}$ denotes the *Fourier transform operator* [Bates and McDonnell, 1989, Sec. 6]. When the far field pattern is evaluated, (2.31) demonstrates that the only components of $\mathbf{E}(\mathbf{r}')$ which are required are those tangential to S.

The Fourier transforms of the components of $\mathbf{E}(x,y)$ tangential to S provide more information than just the observable far field. From (2.28), u and v are only defined for the disk $(u^2+v^2) \leq 1/\lambda^2$. However, the Fourier transforms in (2.31) can be evaluated at all points (u,v) in the range $-\infty < u < \infty$, $-\infty < v < \infty$. When (u,v) is within the disk $(u^2+v^2) \leq 1/\lambda^2$, $\acute{\mathbf{E}}(u,v)$ corresponds to the radiated field. However, when (u,v) is outside this disk, $\acute{\mathbf{E}}(u,v)$ corresponds to the reactive field [Collin and Zucker, 1969a, Sec. 3.2].

The far field pattern is often only required for directions close to $\hat{\mathbf{z}}$. The *small angle region* of the field radiated by the antenna is here defined by

$$\theta \ll 1 \tag{2.32}$$

where θ is the angle, in radians, from the positive z axis. It is apparent from (2.28) that in the small angle region, $w(u,v) \approx 1/\lambda$, $u \ll 1/\lambda$ and $v \ll 1/\lambda$. In the *small angle far field region*, which is the intersection of the small angle and far field regions, (2.31) reduces to

$$\begin{aligned} \acute{E}_x(u,v) &= \frac{je^{-jkR}}{\lambda R} \mathrm{FT}\{E_x(x,y)\} \\ \acute{E}_y(u,v) &= \frac{je^{-jkR}}{\lambda R} \mathrm{FT}\{E_y(x,y)\} \end{aligned} \tag{2.33}$$

while $\acute{E}_z(u,v)$ is negligible compared to the greater of $|\acute{E}_x(u,v)|$ and $|\acute{E}_y(u,v)|$.

2.1.3.3 The aperture field method

The Fourier transform formulation is ideally suited for aperture antennas (Sec. 1.3.3). The plane S can conveniently be taken to be the aperture plane, with S_0 corresponding to the area occupied by the aperture of the antenna. The Fourier transform relationship (2.31) between the aperture and far fields is exact provided that:

1. The tangential components of the aperture field are zero outside a finite region S_0.
2. The radius R, of the hemisphere over which the far field is determined, tends to infinity.
3. The tangential components of the aperture field are known exactly.

The first of these conditions can only be guaranteed if the aperture is a hole in an infinite planar conducting surface (see the boundary condition for a conductor (1.12)). This condition is partially imposed if there is a conducting flange around the aperture. However, by definition, most of the electromagnetic radiation of an aperture antenna passes through its aperture, so the field in the aperture plane is usually negligible outside the aperture.

The requirement for $R \to \infty$ stems from the two conditions stated in (2.22), which assume that $\mathbf{r}$ is parallel to $\mathbf{r}_d$. This holds, effectively, when R is greater than the minimum far field distance, which is defined by (1.20).

Silver [1949, Sec. 5.14] stipulates a further condition on the validity of the Fourier transform relationship between the aperture and far fields, requiring the aperture phase to be almost uniform. Silver derives the Fourier transform relationship from the scalar Kirchhoff integral. This integral is only valid over an open surface, such as S_0, under special conditions (see discussion by Rusch and Potter [1970, Sec. 2.63]), which require a uniform phase distribution. However, the derivation in the previous section is based on a vector integral (2.20), which is always valid over an open surface [Silver, 1949, Sec. 5.11]. Therefore, the phase of both tangential components of the aperture field can be arbitrary, provided the aperture field is consistent with the conditions stated at the beginning of this section.

In practice, the aperture field distribution can never be known exactly. In the *aperture field method* [Silver, 1949, Sec. 5.11], the aperture field is calculated, from the field of the feed and the reflector geometry, by the GO method. Fourier transformation can then be invoked to calculate the far field pattern. For directions close to $\hat{\mathbf{z}}$, the aperture field method yields equivalent results to the physical optics method. But in directions outside the small angle region, the aperture field method is less accurate than the PO method because it does not account for the effects of diffraction between the reflector and the aperture [Rusch and Potter, 1970, Sec. 3.31]. However, due to the ease of calculating Fourier transforms (see Sec. 3.4.1.4), the aperture field method is often simpler to apply than the PO method.

2.1.3.4 Inverse Fourier transformation

Inverse Fourier transformation can be invoked to calculate the tangential aperture field distribution if the far field pattern is known. As pointed out at the end of Section 2.1.2.2, the field $\mathbf{E}(\mathbf{r})$, measured over the radiation hemisphere, has no component in the radial direction. Therefore, if the far field tangential to the hemisphere is measured, its components in the x and y directions can be straightforwardly evaluated.

If the antenna is directional enough for the far field to become insignificant as (u^2+v^2) approaches $1/\lambda^2$, $\acute{\mathrm{E}}(u,v)/w(u,v)$ can be set to zero for $(u^2+v^2) \geq 1/\lambda^2$. Under this condition, the tangential components of the aperture field distribution can be accurately calculated from the measured far field pattern, using (2.31) and the *inverse Fourier transform operator* $\mathrm{IFT}\{\cdot\}$ [Bates and McDonnell, 1989, Sec. 6]:

$$\begin{aligned} E_x(\mathbf{r}') = E_x(x,y) &= \frac{-jR}{e^{-jkR}}\mathrm{IFT}\left\{\frac{\acute{E}_x(u,v)}{w(u,v)}\right\} \\ &= \frac{-jR}{e^{-jkR}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\frac{\acute{E}_x(u,v)}{w(u,v)}\right)e^{-j2\pi(ux+vy)}\,du\,dv \\ E_y(\mathbf{r}') = E_y(x,y) &= \frac{-jR}{e^{-jkR}}\mathrm{IFT}\left\{\frac{\acute{E}_y(u,v)}{w(u,v)}\right\} \\ &= \frac{-jR}{e^{-jkR}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(-\frac{\acute{E}_y(u,v)}{w(u,v)}\right)e^{j2\pi(ux+vy)}\,du\,dv \end{aligned} \tag{2.34}$$

For antennas that are so directional that their far field patterns are significant only in the small angle region (2.32), (2.34) can be simplified to

$$\begin{aligned} E_x(x,y) &= \frac{-jR}{e^{-jkR}}\mathrm{IFT}\left\{\acute{E}_x(u,v)\right\} \\ E_y(x,y) &= \frac{-jR}{e^{-jkR}}\mathrm{IFT}\left\{\acute{E}_y(u,v)\right\} \end{aligned} \tag{2.35}$$

2.1.3.5 Approximations in the Fresnel region

The Fourier transform relationship, expressed by (2.31), is based on approximations (Sec. 2.1.2.2) which are only valid in the far field region. In the near field region, less restrictive approximations must be introduced.

Employing the cosine rule, an exact expression for r_d in (2.18) is

$$\begin{aligned} r_\mathrm{d} &= \left[|\mathbf{r}|^2 + |\mathbf{r}'|^2 - 2\mathbf{r}\cdot\mathbf{r}'\right]^{1/2} \\ &= r\left[1 + \frac{r'^2}{r^2} - 2\frac{r'}{r}\hat{\mathbf{r}}\cdot\hat{\mathbf{r}}'\right]^{1/2} \end{aligned} \tag{2.36}$$

which, when binomially expanded, gives

$$r_\mathrm{d} = r - 2r'(\hat{\mathbf{r}}\cdot\hat{\mathbf{r}}') + \frac{1}{2}\frac{r'^2}{r} - \frac{1}{2}\frac{r'^2}{r}\hat{\mathbf{r}}\cdot\hat{\mathbf{r}}' + \frac{1}{2}\frac{r'^3}{r^2}\left(\hat{\mathbf{r}}\cdot\hat{\mathbf{r}}' - 3(\hat{\mathbf{r}}\cdot\hat{\mathbf{r}}')^3\right) + \ldots \tag{2.37}$$

Equation (2.23) expresses the far field radiated by a source. As stated in (2.22), the phase terms of this expression utilize an approximation for r_d consisting of the first two terms in (2.37). The error in this approximation is given by the remaining terms in (2.37) and can never be greater than $r'^2/(2r)$. At the minimum far field distance, defined by (1.20), this error has a maximum value of $\lambda/16$, when the maximum value of r' is half the largest dimension of the source.

The *Fresnel region* is defined to be that region in which, when considering the factor e^{-jkr_d} in (2.21), the first four terms in (2.37) must be retained in the approximation to r_d [Collin and Zucker, 1969a, p. 40]. Although the boundaries of the Fresnel region are not clearly delimited, it can be considered to be that part of the near field region for which $r \gg r'_{\mathrm{max}}$, where r'_{max} is the largest value of r' [Collin and Zucker, 1969a, p. 40].

Despite this, it is convenient to define what is here called the *Fourier Fresnel region*, within which (cf. (2.22))

$$\begin{aligned} r_{\mathrm{d}} &= \begin{cases} r & \text{except in phase expressions} \\ r - \mathbf{r}' \cdot \hat{\mathbf{r}} - r'^2/(2r) & \text{in phase expressions} \end{cases} \\ \hat{\mathbf{r}}_{\mathrm{d}} &= \hat{\mathbf{r}} \end{aligned} \tag{2.38}$$

Note that the approximations involved in the far field and Fourier Fresnel regions differ only as regards the phase terms in (2.21). For amplitude terms, it is assumed in both regions that $\mathbf{r}$ and $\mathbf{r}_{\mathrm{d}}$ are approximately equal in magnitude and direction.

When (2.38) is used in place of (2.22), a similar development to that presented in Section 2.1.3.2 leads to

$$\begin{aligned} \grave{E}_x(u,v) &= jw(u,v)\frac{e^{-jkR}}{R}\mathrm{FT}\Big\{E_x(x,y)e^{-jk(x^2+y^2)/(2R)}\Big\} \\ \grave{E}_y(u,v) &= jw(u,v)\frac{e^{-jkR}}{R}\mathrm{FT}\Big\{E_y(x,y)e^{-jk(x^2+y^2)/(2R)}\Big\} \\ \grave{E}_z(u,v) &= \frac{-u}{w(u,v)}\grave{E}_x(u,v) + \frac{-v}{w(u,v)}\grave{E}_y(u,v) \end{aligned} \tag{2.39}$$

instead of (2.33). The radius R is here less than the minimum far field distance and $\grave{\mathbf{E}}(u,v)$ denotes the *Fourier Fresnel pattern*.

Comparing (2.37) with (2.38) reveals that the maximum error in the Fourier Fresnel approximation to r_{d} is $r'^2(\hat{\mathbf{r}} \cdot \hat{\mathbf{r}}')/(2r)$. This is at most $\lambda/16$ when

$$\sin^2\theta \leq \frac{r}{R_{\mathrm{ff}}} \tag{2.40}$$

where R_{ff} is the minimum far field distance defined by (1.20) and θ is the angle from the z axis. The Fourier Fresnel region can thus be considered to be the part of the Fresnel region in which (2.40) holds. Therefore, in the Fourier Fresnel and far field regions, the maximum errors in the phase terms, of (2.39) and (2.31) respectively, are acceptably equal for most practical applications.

The *small angle Fresnel region* is defined to be the intersection of the small angle region defined by (2.32) and the Fourier Fresnel region. Applying the approximations for the small angle region (introduced in Sec. 2.1.3.2) to (2.39), the radiation pattern in the small angle Fresnel region is given by

$$\begin{aligned} \grave{E}_x(u,v) &= \frac{je^{-jkR}}{\lambda R}\mathrm{FT}\Big\{E_x(x,y)e^{-jk(x^2+y^2)/(2R)}\Big\} \\ \grave{E}_y(u,v) &= \frac{je^{-jkR}}{\lambda R}\mathrm{FT}\Big\{E_y(x,y)e^{-jk(x^2+y^2)/(2R)}\Big\} \end{aligned} \tag{2.41}$$

2.2 PERFORMANCE

The purpose of a high gain antenna is to exhibit a gain pattern (Sec. 1.2.2) which is mainly concentrated over a small solid angle (usually centred on the direction perpendicular to the aperture plane) and is as small as possible over all other angles. This means that when transmitting, most of the available power is radiated in the desired direction and little is lost to other directions. When receiving, radiation from the desired direction is detected with minimum interference from radiation from other directions.

In addition to the general electrical properties of antennas, discussed in Section 1.2, high gain antennas have further properties which pertain to their far field patterns. By inverse Fourier transformation (Sec. 2.1.3.4) many of these properties can be related to those of the aperture field distribution. Parameters which characterize the properties of high gain antennas are discussed in the following sections.

2.2.1 Features of a gain pattern

Figure 2.7 illustrates the general appearance of a cross-section through a gain pattern of a high gain antenna. The unit of the vertical axis is the *decibel (dB)*, which expresses a ratio between two power levels [IEEE, 1984]. From (1.21), the gain is defined as a power ratio and therefore can be expressed in decibels as $10 \log_{10} G(\theta; \phi)$.

A gain pattern typically comprises many lobes. The *main beam* is by definition the largest of these lobes. The *boresight* [IEEE, 1984] is defined to be the direction of the peak of the main beam and is usually perpendicular to the aperture plane of the antenna. The pattern then alternates between *nulls*, which are local minima, and the remaining lobes, which are called *sidelobes*. For a circularly symmetric aperture field distribution, the gain pattern is also circularly symmetric, so the sidelobes and nulls form concentric rings around the circular main beam. The sidelobe level tends to fall off with angle from boresight. Wide angle sidelobes of appreciable magnitude are often exhibited by reflector antennas. The field can even build up in directions behind a reflector antenna when currents running in the reflector's edges are appropriately phased [Collin and Zucker, 1969b, p. 49].

The *half power beamwidth* of an antenna, taken in a plane containing the boresight, is the angle between the two directions in which the radiated power is one half (i.e. 3 dB down on) that of the maximum radiated power [IEEE, 1984].

2.2.2 Relationship between gain pattern and aperture field distribution

Equation (2.31) shows the intimate relationship between the aperture and far fields. This suggests that the aperture field distribution has a direct influence on an antenna's gain pattern.

The definition of a gain pattern (1.21) can be rearranged to separate the contribution of the antenna's radiation efficiency η_{rad} (Sec. 1.2.4) from the directional characteristics of the radiation pattern, i.e.

$$G(\theta; \phi) = \eta_{\text{rad}} \frac{4\pi P_{\text{rad}}(\theta; \phi)}{\int_0^{2\pi} \int_0^{\pi} P_{\text{rad}}(\theta; \phi) \sin\theta \, d\theta \, d\phi} \tag{2.42}$$

where $P_{\text{rad}}(\theta; \phi)$ is the radiated power per unit solid angle and the denominator is equal to the total radiated power. Since the far field behaves locally as a plane wave, the Poynting vector can be derived from the far field via (2.3). From Poynting's theorem (Sec. 1.1.5) $P_{\text{rad}}(\theta; \phi)$ can be defined as

$$P_{\text{rad}}(\theta; \phi) = \frac{1}{2} \left(\frac{\epsilon}{\mu} \right)^{1/2} \left(|E_x(\theta; \phi)|^2 + |E_y(\theta; \phi)|^2 + |E_z(\theta; \phi)|^2 \right) R^2 \tag{2.43}$$

where $|E_x(\theta; \phi)|$, $|E_y(\theta; \phi)|$ and $|E_z(\theta; \phi)|$ are amplitude patterns (Sec. 1.2.2) of the three Cartesian components of the far field on a sphere of radius R.

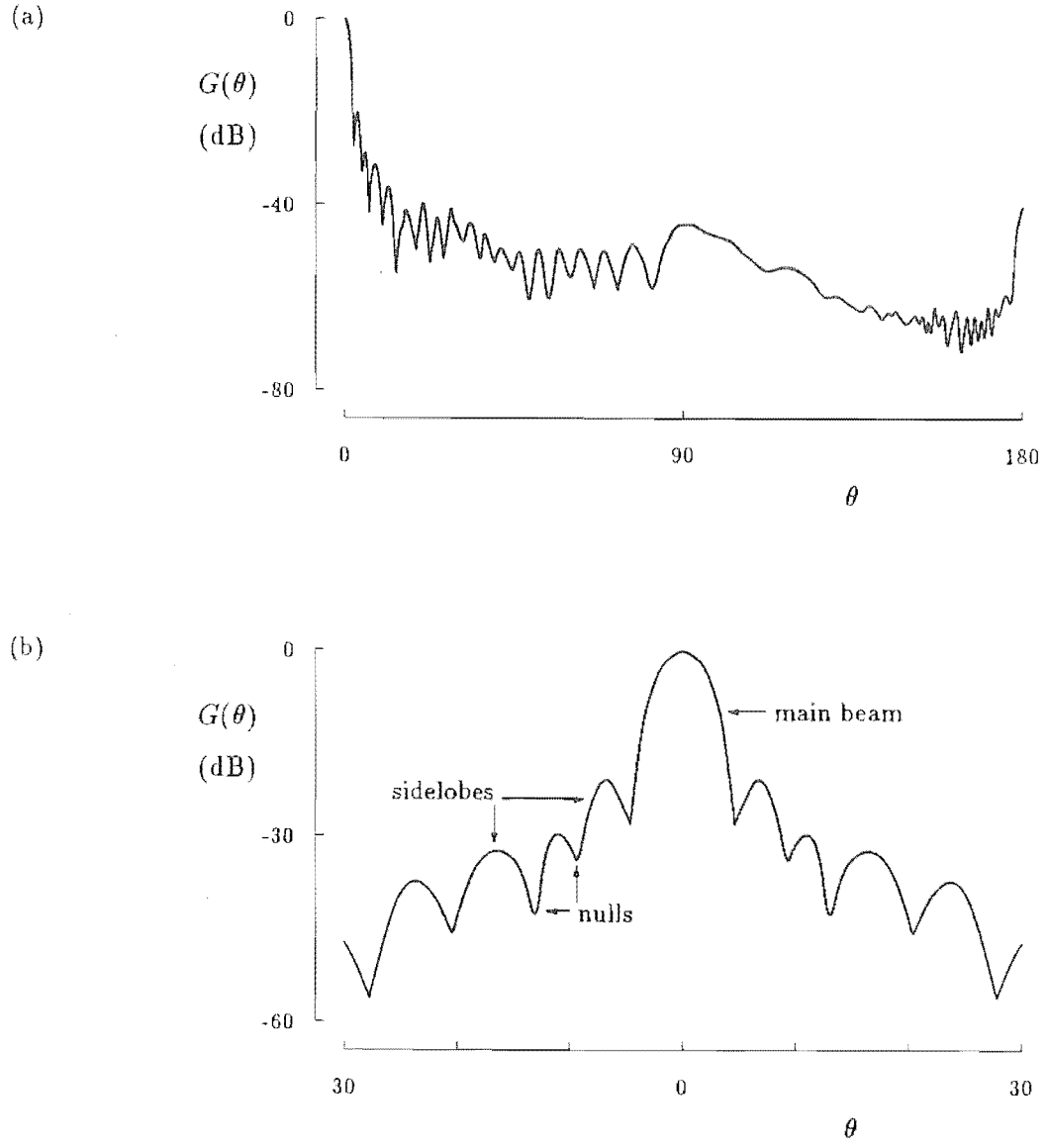


Figure 2.7 A cut through the gain pattern of a high gain antenna, showing some of its features: (a) the pattern from the front to the back of the antenna; (b) an expanded view of the region around the main beam. The angle from boresight is θ. Loosely based on an antenna analysis by James [1980].

In accord with the aperture field method (Sec. 2.1.3.3), the part of the field radiated in directions behind the aperture plane is neglected. Substituting (2.43) into (2.42) and converting to the u, v coordinates (see (2.28)) yields

$$G(u,v) = \eta_{\text{rad}} \frac{4\pi \left(|\acute{E}_x(u,v)|^2 + |\acute{E}_y(u,v)|^2 + |\acute{E}_z(u,v)|^2 \right)}{\iint_{S_{1/\lambda}} \frac{\lambda}{w(u,v)} \left(|\acute{E}_x(u,v)|^2 + |\acute{E}_y(u,v)|^2 + |\acute{E}_z(u,v)|^2 \right) du\, dv} \tag{2.44}$$

where $S_{1/\lambda}$ is the disk in the u, v plane, for which $(u^2 + v^2) \leq (1/\lambda)^2$, corresponding to the radiation hemisphere. Because of the way (2.44) has been normalized, only the relative amplitude patterns are required for the calculation of the gain pattern.

In order to simplify (2.44), consider an aperture field which is linearly polarized in the x direction (i.e. $E_y(x,y) = 0$). It is assumed that the antenna is so highly directional that the radiated field is only significant within the small angle region (2.32). Appealing to (2.33), (2.44) is seen to approximate to

$$G(u,v) = \eta_{\text{rad}} \frac{4\pi |\text{FT}\{E_x(x,y)\}|^2}{\lambda^2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |\text{FT}\{E_x(x,y)\}|^2 \, du\, dv} \tag{2.45}$$

In this equation, $E_x(x,y)$ can be expressed relative to any constant field value without affecting the result.

Assuming that the peak gain occurs in the z direction ($u = v = 0$) and employing the energy conservation theorem [Bates and McDonnell, 1989, p. 24], the peak gain is given by [Collin and Zucker, 1969a, p. 79]

$$G_{\text{max}} = \eta_{\text{rad}} \frac{4\pi \left| \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} E_x(x,y)\, dx\, dy \right|^2}{\lambda^2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |E_x(x,y)|^2 \, dx\, dy} \tag{2.46}$$

2.2.3 Uniform aperture field distribution

Subject to the conditions required to derive (2.46), it is possible to compare the peak gains produced by different kinds of aperture field distributions. Note that the aperture field distribution here consists of only a single vector component, so it can be treated as a scalar distribution. Silver [1949, Sec. 6.4] shows that, for a given sized aperture, the aperture field distribution which produces the maximum gain is the *uniform distribution*. This is an aperture field distribution in which the amplitude and phase are constant within the aperture (and the amplitude is zero outside the aperture). The peak gain produced by a uniform distribution is

$$G_{\text{max}} = \eta_{\text{rad}} \frac{4\pi A}{\lambda^2} \tag{2.47}$$

where A is the area of the aperture.

The half power beamwidth of a circular aperture with a uniform field distribution is [Silver, 1949, Sec. 6.8]

$$\theta_{\text{bw}} = 2 \sin^{-1} \left(0.51 \frac{\lambda}{D} \right) \approx 1.02 \frac{\lambda}{D} \tag{2.48}$$

where D is the diameter of the aperture. It can be seen that the larger the aperture, relative to the wavelength, the higher the peak gain and the narrower the main beam.

2.2.4 Aperture efficiency

When the aperture field distribution is not uniform, the peak gain is less than that given in (2.47). A measure of this is the *aperture efficiency* η_{aper} which, for a lossless antenna, is defined by [IEEE, 1984, p. 46]

$$\eta_{\text{aper}} = \frac{(A_e)_{\max}}{A} \tag{2.49}$$

where $(A_e)_{\max}$ is the maximum effective area of the antenna (see (1.22)) and A is the physical aperture area. From the relationship between gain and effective area (1.23), the peak gain of a (lossy) antenna is given by

$$G_{\max} = \eta_{\text{aper}}\eta_{\text{rad}}\frac{4\pi A}{\lambda^2} \tag{2.50}$$

Comparing this equation with (2.47), η_{aper} is seen to be a measure of the the efficiency of an antenna's ability to concentrate its radiated power in the neighbourhood of a desired direction. For an antenna which is receiving rather than transmitting radiation, (2.49) shows that η_{aper} is a measure of the ability of a given aperture to collect power incident upon it.

2.2.5 Non-uniform aperture field distributions

For a given aperture field amplitude distribution $|E_x(x,y)|$, the peak gain is maximum when the numerator of (2.46) is maximum. This occurs when the phase of the aperture field is uniform.

Any defects of the antenna geometry affect the phase distribution of the aperture field (see Sec. 3.2). Because reflectors can only be manufactured to a finite tolerance, they inevitably cause random fluctuations in the phase of the aperture field. Ruze [1966] has considered the effect on the gain pattern of such a random aperture field phase distribution. It is assumed that the phase at (x,y) in the aperture belongs to a Gaussian population with a mean of zero and a standard deviation of σ. Furthermore, the degree of correlation between the phase at two points separated by a distance ρ is assumed to be $\exp(-\rho^2/c^2)$, where c is called the correlation radius. It is also assumed that c is constant over the the aperture and is much less than the aperture diameter. It is also assumed that the amplitude of the aperture field is approximately constant over distances of the order of c [Rudge *et al.*, 1982, Sec. 3.1.2]. Employing a physical optics model, Ruze [1966] obtains the following formulation for the statistically expected gain pattern $G(u,v)$:

$$G(u,v) = G_0(u,v)e^{-\sigma^2} + \left(\frac{2\pi c}{\lambda}\right)^2 e^{-\sigma^2}\sum_{n=1}^{\infty}\frac{\sigma^{2n}}{n!\,n}e^{(\pi c)^2(u^2+v^2)/n} \tag{2.51}$$

where $G_0(u,v)$ is the gain when the aperture field phase distribution is uniform. At the centre of the pattern, for reasonable values of σ and c, the second term in (2.51) can be neglected, yielding a boresight gain of

$$G(0,0) = G_0(0,0)e^{-\sigma^2} \tag{2.52}$$

The above two equations indicate that the effect of a random phase distribution is to remove some of the power radiated in the boresight direction, reradiating it over a wide range of angles. An example of the gain pattern for a particular random aperture field

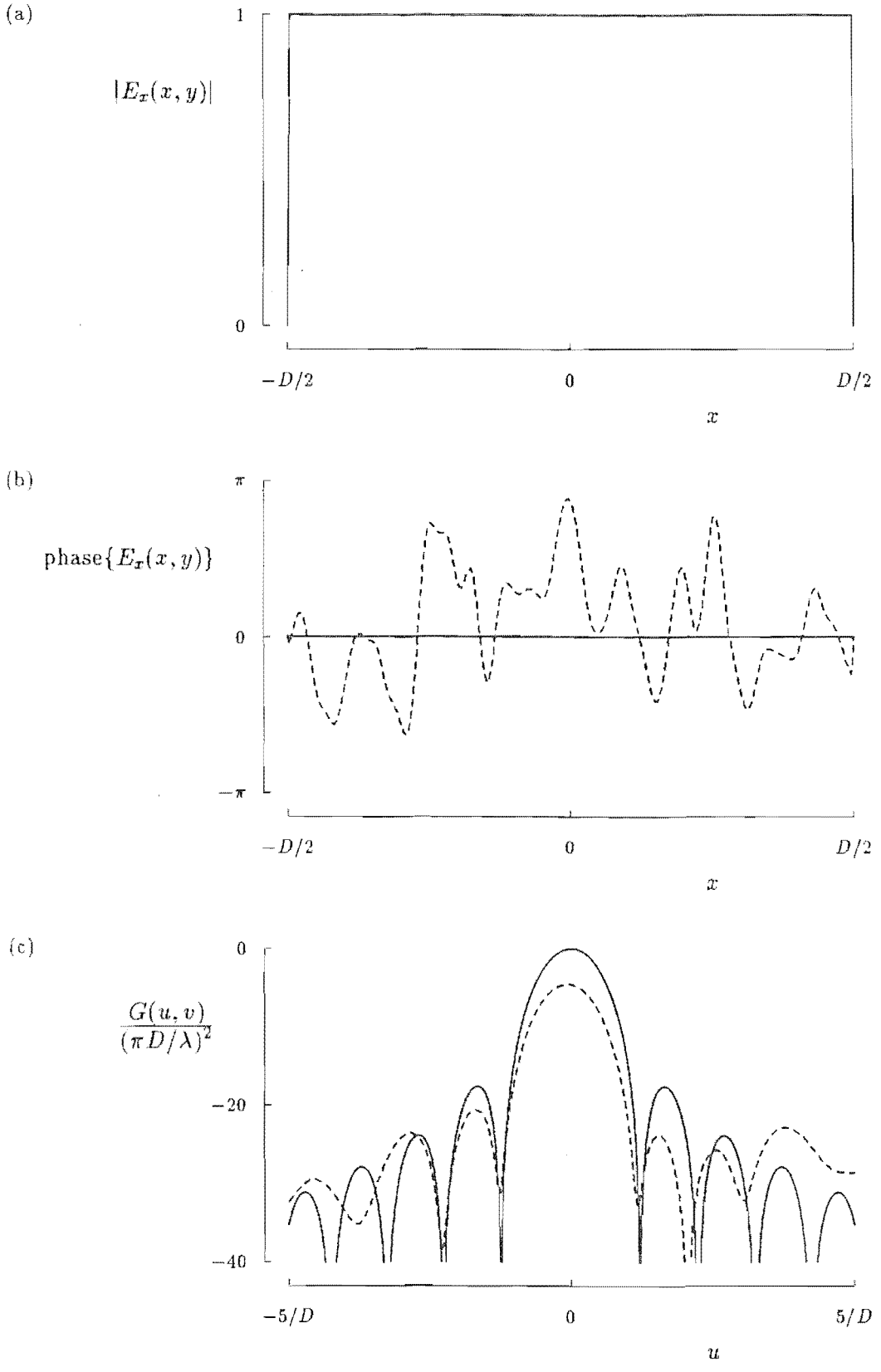


Figure 2.8 Comparison of the gain patterns produced by a uniform (solid curve) and a random (dashed curve) aperture field phase distribution: (a) aperture field amplitude distributions; (b) aperture field phase distributions; (c) gain patterns. Both aperture field distributions have an amplitude of unity inside, and zero outside, a disk of diameter D. The random phase distribution has a standard deviation of $\sigma = 1$ and a correlation radius of $c = 0.02D$.

phase distribution is shown in Figure 2.8. As with all the examples presented in this section, (2.45) is invoked to derive the gain pattern from the aperture field distribution.

A different kind of aperture field phase distribution is one which is highly correlated over the whole aperture. An example of such a distribution is the radially quadratic aperture field phase distribution produced by a defocused antenna (see Sec. 3.2.2). An example of this is shown in Figure 2.9. The aperture field phase distributions shown in Figures 2.9 and 2.8 have equal rms values. A comparison of these figures confirms the trend predicted by Davis [1970, p. 38]: with large correlation distances the main beam broadens and the sidelobe structure is generally less distinct, while for small correlation distances the shape of the gain pattern near to boresight is relatively unchanged, with the sidelobes far from boresight becoming less distinct. For both of these aperture field phase distributions, the peak gain is reduced and the sidelobe levels are increased.

In principle, the phase distribution of the aperture field can be kept adequately uniform by constructing the antenna to within a sufficiently tight tolerance. To provide maximum peak gain, not only must the phase of the aperture field be uniform, but also the power from the antenna feed must be distributed uniformly over the aperture. However, with most feeds, this results in significant *spillover* [Rudge *et al.*, 1982, Sec. 3.2.2], which is power radiated from the feed that does not illuminate the reflector, but instead radiates past it. Because the power is not reflected towards the aperture, it in general does not contribute to the peak gain of the antenna, and therefore reduces the aperture efficiency.

Spillover can be reduced by allowing a *tapered distribution* of the aperture field amplitude. In antenna design there is a tradeoff between the losses of aperture efficiency due to spillover and nonuniformity of the aperture field distribution [Wood, 1980, p. 3].

Figure 2.10 compares the calculated gain patterns produced by a uniform and a particular tapered aperture amplitude distribution. The relatively small loss in gain due to tapering is evident. However, the tapered distribution produces sidelobe levels that are much smaller than those due to the uniform distribution. Note that reduction in sidelobe levels can only be achieved at the expense of reduction of peak gain and broadening of the main beam.

Another cause of nonuniformity of the aperture field amplitude is *aperture blockage*. This occurs when objects (such as struts, or the feed) lie in the path of the rays from the feed to the aperture. The geometrical optics method predicts that the aperture field is zero in the geometrical shadows of the objects. More accurate analysis (based, for instance, on GTD — see Sec. 2.1.1.5) predicts a non-zero field in these regions due to diffraction effects. Figure 2.11 illustrates the effects on the gain pattern of aperture blockage. In general, the peak gain decreases, the main beam broadens and the levels of some of the sidelobes increase. These effects become more pronounced as the area of the blockage shadow relative to the aperture area increases.

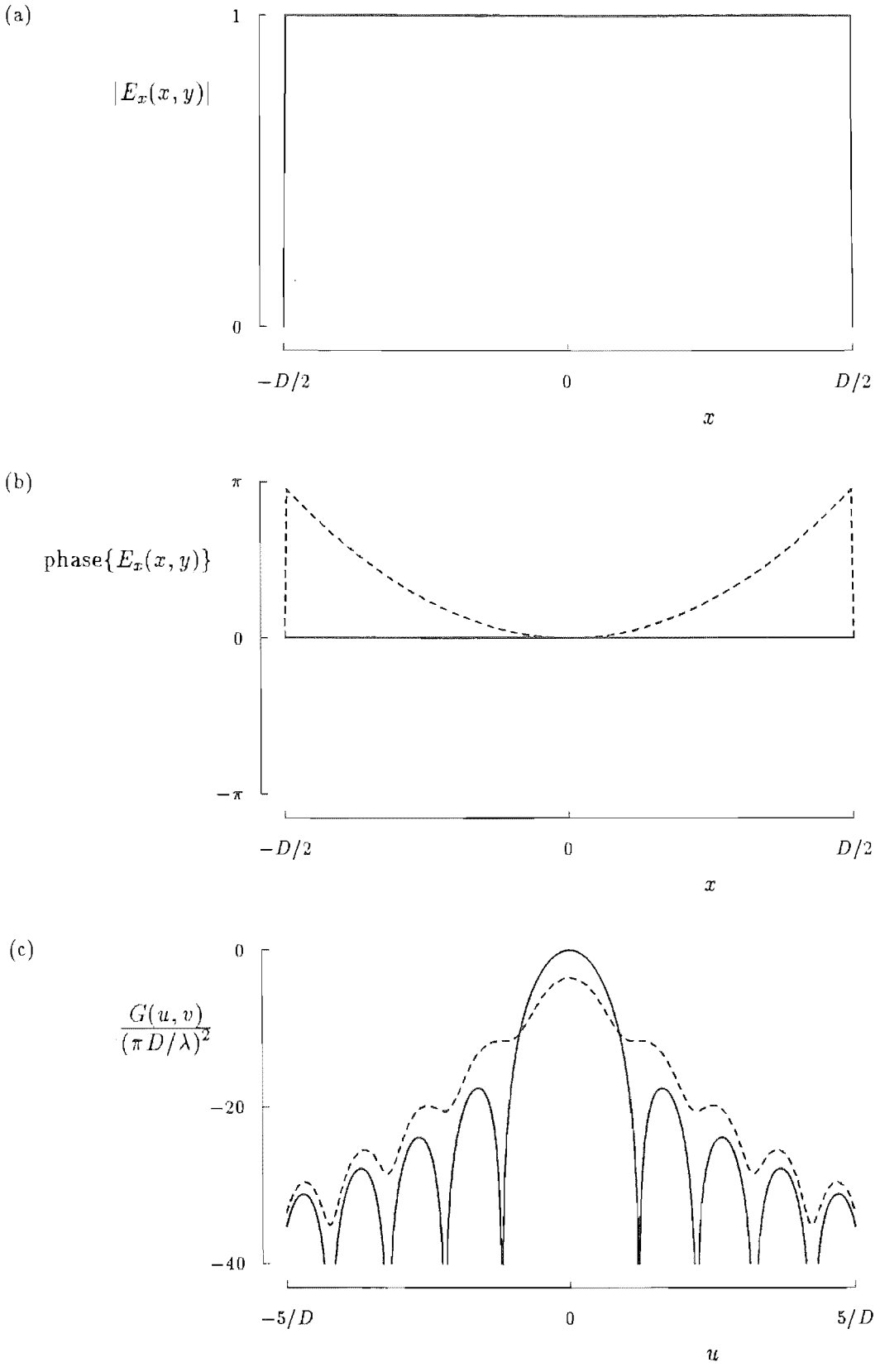


Figure 2.9 Comparison of the gain patterns produced by a uniform (solid curve) and a quadratic (dashed curve) aperture field phase distribution: (a) aperture field amplitude distributions; (b) aperture field phase distributions; (c) gain patterns. Both aperture field distributions have an amplitude of unity inside, and zero outside, a disk of diameter D. The quadratic phase distribution is described by phase$\{E_x(\rho)\} = 3\rho^2$ where $\rho = 2(x^2 + y^2)^{1/2}/D$.

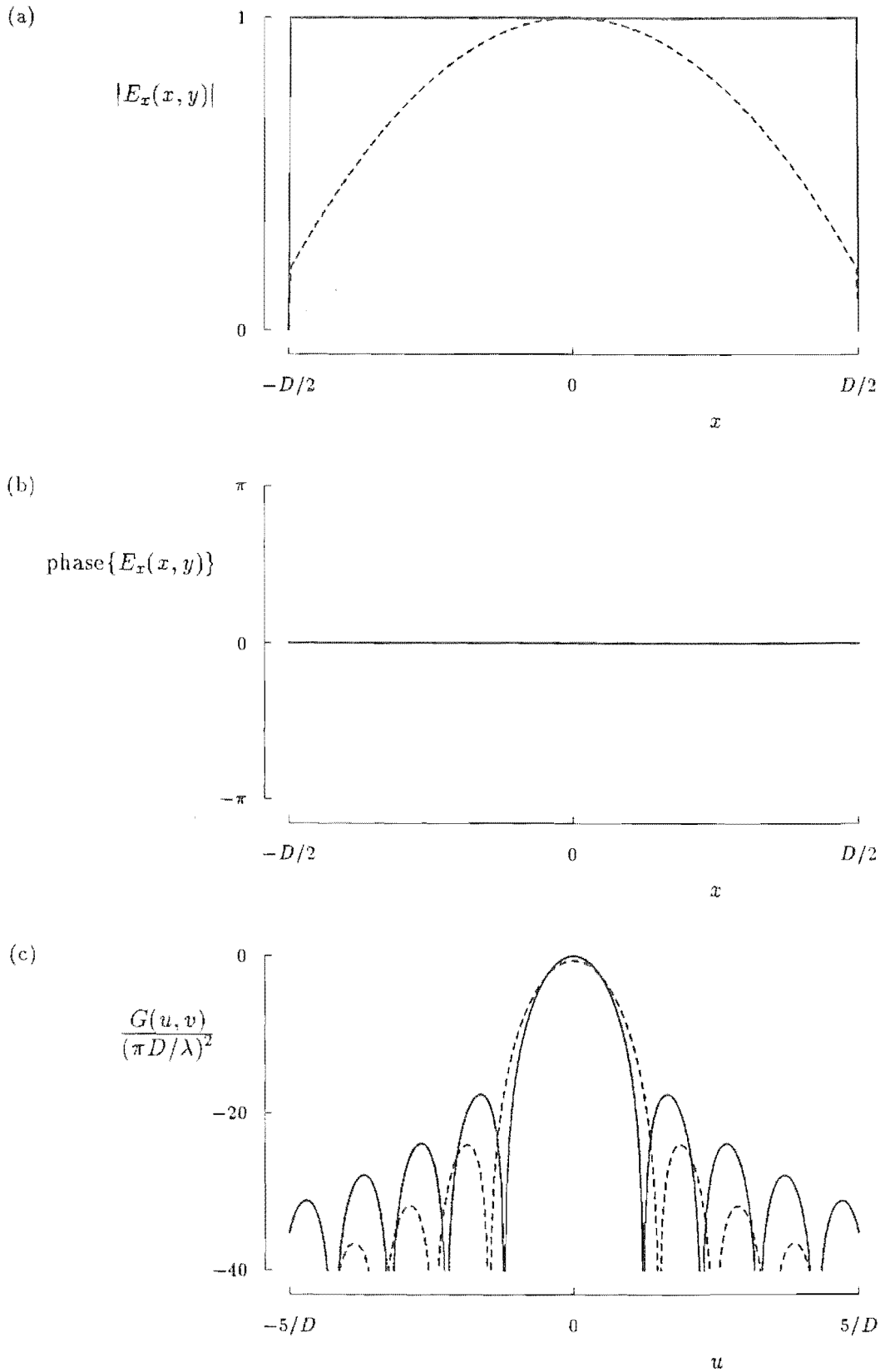


Figure 2.10 Comparison of the gain patterns produced by a uniform (solid curve) and a tapered (dashed curve) aperture field amplitude distribution: (a) aperture field amplitude distributions; (b) aperture field phase distributions; (c) gain patterns. Both aperture distributions are zero outside a disk of diameter D. The uniform distribution is $E_x(x,y) = 1$. The tapered distribution is $E_x(\rho) = 1 - 0.8\rho^2$ where $\rho = 2(x^2 + y^2)^{1/2}/D$.

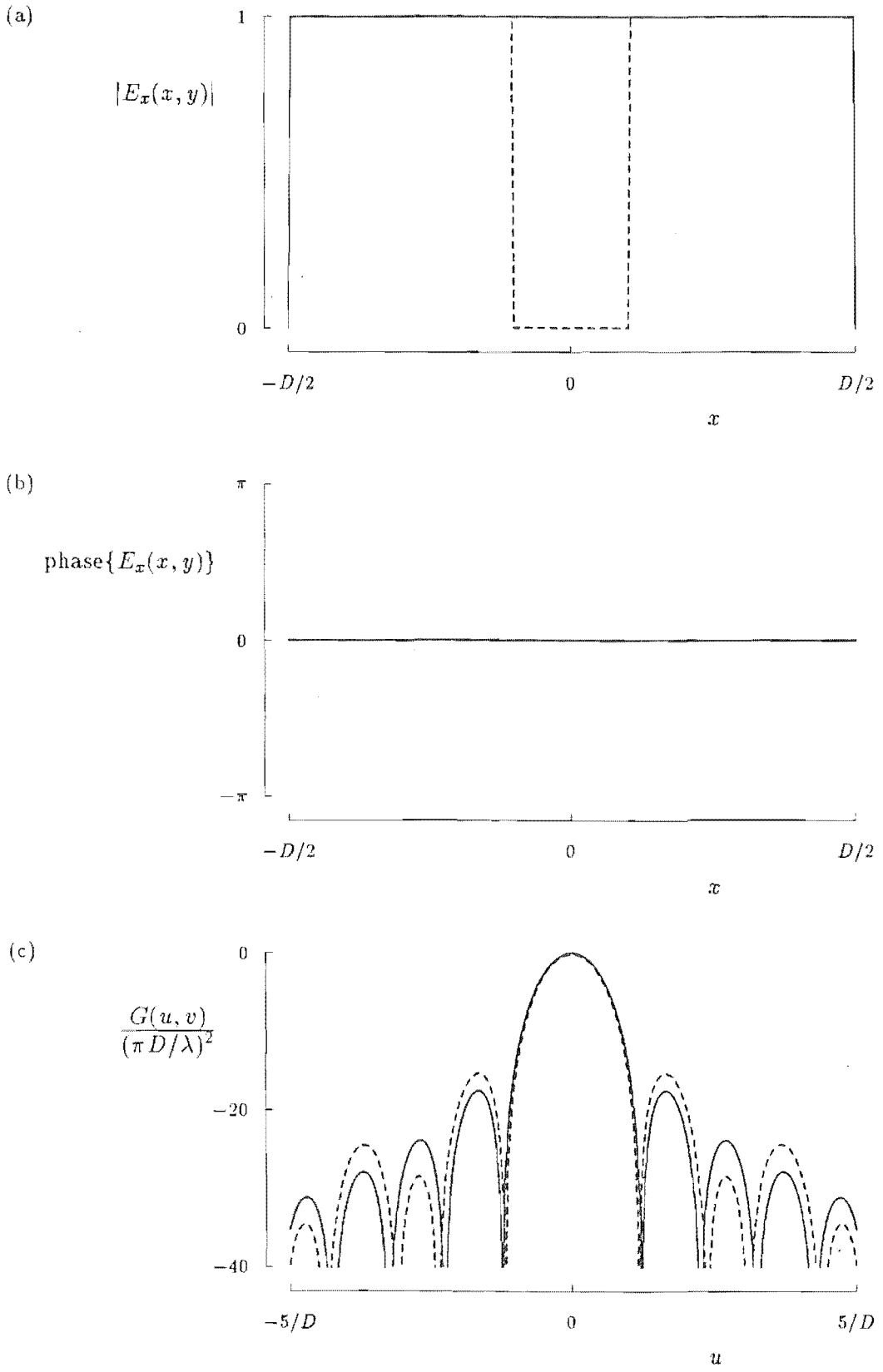


Figure 2.11 Comparison of the gain patterns produced by a uniform aperture field distribution with (solid curve) and without (dashed curve) aperture blockage: (a) aperture field amplitude distributions; (b) aperture field phase distributions; (c) gain patterns. Both field distributions are zero outside a disk of diameter D. The aperture field distribution without blockage is $E_x(x,y) = 1$. The distribution with blockage is $E_x(\rho) = 1$ for $\rho > 0.1$ and $E_x(\rho) = 0$ for $\rho < 0.1$, where $\rho = 2(x^2 + y^2)^{1/2}/D$.

2.2.6 Polarization

In a point to point communications system, a transmitting antenna radiates as much power as possible towards the receiving antenna. However, in order for the receiving antenna to extract maximum power from the wave, the wave must be polarized in a manner compatible with the polarization characteristics of the antenna. For example, a vertically polarized receiving antenna does not extract any power from a horizontally polarized incident wave, but extracts maximum power from a vertically polarized wave. This concept of polarization matching is discussed in a general way in this section.

For a transmitting antenna, the angular variation of the polarization unit vector $\hat{\imath}_{\text{rad}}(\theta;\phi)$ of the radiated field can be readily deduced from the antenna's far field pattern (Sec. 1.2.2), using (1.17). Consider the same antenna receiving an incident uniform plane wave (Sec. 2.1.1.1) from a direction $(\theta_0;\phi_0)$. The Poynting vector of this wave is $\mathbf{P}_{\text{inc}}$ and its polarization unit vector is $\hat{\imath}_{\text{inc}}$. From reciprocity considerations, Yeh [1949] shows that the power P_{out}, transferred from the antenna to a matched load, is

$$P_{\text{out}} = \frac{\lambda^2}{4\pi} |\mathbf{P}_{\text{inc}}| \, G(\theta_0;\phi_0) \, | \, \hat{\imath}_{\text{inc}} \cdot \hat{\imath}_{\text{rad}}(\theta_0;\phi_0)|^2 \tag{2.53}$$

where $G(\theta;\phi)$ is the gain pattern of the antenna.

The *polarization efficiency* of an antenna, for a particular incoming plane wave, is the ratio of power received by the load to the power the load would have received had the wave's polarization been adjusted for maximum power reception [IEEE, 1984]. From (2.53) the polarization efficiency η_{pol} is given by

$$\eta_{\text{pol}} = | \, \hat{\imath}_{\text{inc}} \cdot \hat{\imath}_{\text{rad}} \, |^2 \tag{2.54}$$

The maximum received power occurs when the incident field is *polarization matched* to the far field pattern, that is, when $\hat{\imath}_{\text{inc}} = \hat{\imath}^*_{\text{rad}}$.

In practice, antennas are often required to transmit or receive a particular polarization, called the *copolarization*. The copolarization is taken as the reference polarization and must be specified for each direction $(\theta;\phi)$ [e.g. Ludwig, 1973]. The *cross polarization* is defined to be everywhere orthogonal to the copolarization, in the same plane as the copolarization. The conversion of power intended to be copolarized into cross polarization is called *depolarization*.

So far in this thesis, the far field pattern $\mathbf{E}(\theta;\phi)$ of an antenna has been expressed in terms of its Cartesian components E_x, E_y, E_z. However, it can also be expressed in terms of its copolar and cross polar components, E_{co} and E_{x} respectively, and a radial component which is zero. Section 1.1.6 describes how to calculate the vector component of a field corresponding to any given polarization unit vector. Substituting (1.17) into (2.54), the polarization efficiency of an antenna, for a copolarized plane wave incident from a direction $(\theta_0;\phi_0)$, is [IEEE, 1979, Sec. 11.1]

$$\begin{aligned} \eta_{\text{pol}} &= \left| \frac{\mathbf{E}(\theta_0;\phi_0)}{|\mathbf{E}(\theta_0;\phi_0)|} \cdot \hat{\imath}^*_{\text{co}}(\theta_0;\phi_0) \right|^2 \\ &= \frac{|E_{\text{co}}(\theta_0;\phi_0)|^2}{|E_{\text{co}}(\theta_0;\phi_0)|^2 + |E_{\text{x}}(\theta_0;\phi_0)|^2} \end{aligned} \tag{2.55}$$

where $|E_{\text{co}}(\theta_0;\phi_0)|$ and $|E_{\text{x}}(\theta_0;\phi_0)|$ are the copolar and cross polar amplitude patterns respectively. For a transmitting antenna, η_{pol} is a measure of the proportion of the total power, transmitted in direction $(\theta_0;\phi_0)$, which is copolarized.

The orientation of the copolarization unit vector must always vary with position over the radiation hemisphere, because by definition it must always be perpendicular to the direction of propagation. However, in the small angle far field region (Sec. 2.1.3.2), the direction of propagation is approximately constant (and parallel to $\hat{\mathbf{z}}$). Therefore $\hat{\imath}_{\rm co}$ and $\hat{\imath}_{\rm x}$ can be defined as constants in the form

$$\hat{\imath}_{\rm co} = a\,\hat{\mathbf{x}} + b\,\hat{\mathbf{y}} \quad \text{and} \quad \hat{\imath}_{\rm x} = b^*\,\hat{\mathbf{x}} - a^*\,\hat{\mathbf{y}} \tag{2.56}$$

where a and b are complex constant scalars which satisfy $(|a|^2 + |b|^2) = 1$. These same polarization unit vectors can describe the tangential components of the aperture field. From (1.18) and invoking the linearity property of Fourier transformation, (2.33) yields

$$\begin{aligned} \acute{E}_{\rm co}(u,v) &= \frac{je^{-jkR}}{\lambda R}\mathrm{FT}\{E_{\rm co}(x,y)\} \\ \acute{E}_{\rm x}(u,v) &= \frac{je^{-jkR}}{\lambda R}\mathrm{FT}\{E_{\rm x}(x,y)\} \end{aligned} \tag{2.57}$$

where $E_{\rm co}(x,y)$ and $E_{\rm x}(x,y)$ are the *copolar and cross polar aperture field distributions* respectively, and $\acute{E}_{\rm co}(u,v)$ and $\acute{E}_{\rm x}(u,v)$ are the *copolar and cross polar far field patterns* respectively. Equation (2.57) shows that copolar components of the aperture field radiate only copolar field components in the small angle far field region. To produce a radiated field which has no cross polar component in this region, there must be no cross polar component in the aperture field. However, if it is only required that there is to be no cross polar component radiated in the $\hat{\mathbf{z}}$ direction, it is sufficient for the average value of the cross polar field over the aperture to be zero.

An equation similar to (2.57) also holds in the small angle Fresnel region. Combining (2.56), (1.18) and (2.39) yields

$$\begin{aligned} \grave{E}_{\rm co}(u,v) &= \frac{je^{-jkR}}{\lambda R}\mathrm{FT}\left\{E_{\rm co}(x,y)e^{-jk(x^2+y^2)/(2R)}\right\} \\ \grave{E}_{\rm x}(u,v) &= \frac{je^{-jkR}}{\lambda R}\mathrm{FT}\left\{E_{\rm x}(x,y)e^{-jk(x^2+y^2)/(2R)}\right\} \end{aligned} \tag{2.58}$$

where $\grave{E}_{\rm co}(u,v)$ and $\grave{E}_{\rm x}(u,v)$ are the *copolar and cross polar Fourier Fresnel patterns* respectively.

2.2.7 Figure of merit (G/T)

Consider a uniform plane wave, having a Poynting vector $\mathbf{P}_{\rm inc}$, incident from boresight upon a polarization matched antenna. It follows from (2.53) and (1.25) that the ratio of the (wanted) power $P_{\rm out}$, received from the wave, to the (unwanted) power N, received from noise sources, is

$$\frac{P_{\rm out}}{N} = \frac{(\lambda^2/4\pi)\,|\mathbf{P}_{\rm inc}|\,G_{\rm max}}{k\,T_{\rm sys}\,\Delta f} \tag{2.59}$$

where $T_{\rm sys}$ is the system noise temperature, which equals the antenna noise temperature (Sec. 1.2.6) plus the receiving circuitry noise temperature referred to the antenna terminals. For a given incoming field, this ratio is proportional to the *figure of merit* (G/T) of the antenna [IEEE, 1984]:

$$G/T = \frac{G_{\rm max}}{T_{\rm sys}} \tag{2.60}$$

The figure of merit of an antenna is therefore a parameter indicative of the signal to noise ratio of the receiving system.

2.3 CONFIGURATIONS

In practice, there are many ways in which a feed and one or more reflectors can be configured to produce a desired radiation pattern. The following sections discuss the geometry and relative merits of paraboloidal (Sec. 2.3.1), Cassegrain (Sec. 2.3.2) and offset (Sec. 2.3.3) reflector antennas.

2.3.1 Paraboloidal reflectors

A *paraboloidal reflector antenna* consists of a feed and a reflector whose surface is an axially symmetric section of a paraboloid. It is the simplest kind of high gain reflector antenna. Section 2.3.1.1 provides a detailed ray tracing analysis (see Sec. 2.1.1.4) of the paraboloidal reflector, while Section 2.3.1.2 discusses the feeds that can be employed. General comments about paraboloidal reflector antennas are made in Section 2.3.1.3

2.3.1.1 Ray tracing analysis

The aim of this section is to predict the aperture field distribution of a paraboloidal reflector, when it is illuminated by a feed whose radiation characteristics are known. This analysis employs both the Cartesian and the spherical coordinate systems so that any point in space can be expressed as either (x, y, z) or $(r; \theta; \phi)$ (see (1.8)).

Figure 2.12 shows a paraboloid, whose axis is in the z direction and whose focus is at the origin. Its shape is described by the set of points satisfying either

$$\begin{aligned} z &= \frac{x^2 + y^2}{4f} - f \\ \text{or} \quad r &= \frac{2f}{1 - \cos\theta} \end{aligned} \tag{2.61}$$

where f is the focal length of the paraboloid. The two equations of (2.61) are equivalent descriptions of the paraboloid, because substitution of (1.8) into the first equation yields the second equation.

The geometrical path of a ray is determined with the aid of the laws of reflection. Consider a feed, positioned so that it radiates rays which appear to emanate from the origin. This assumes that the reflector is in the far field region of a feed whose phase centre is at the origin. The direction of a ray, incident upon the reflector at point $B = (x, y, z)$, is given by

$$\hat{\mathbf{s}}_i = \frac{x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}} + z\,\hat{\mathbf{z}}}{r} \tag{2.62}$$

From (2.61), the normal to the paraboloid, at a point B, is given by

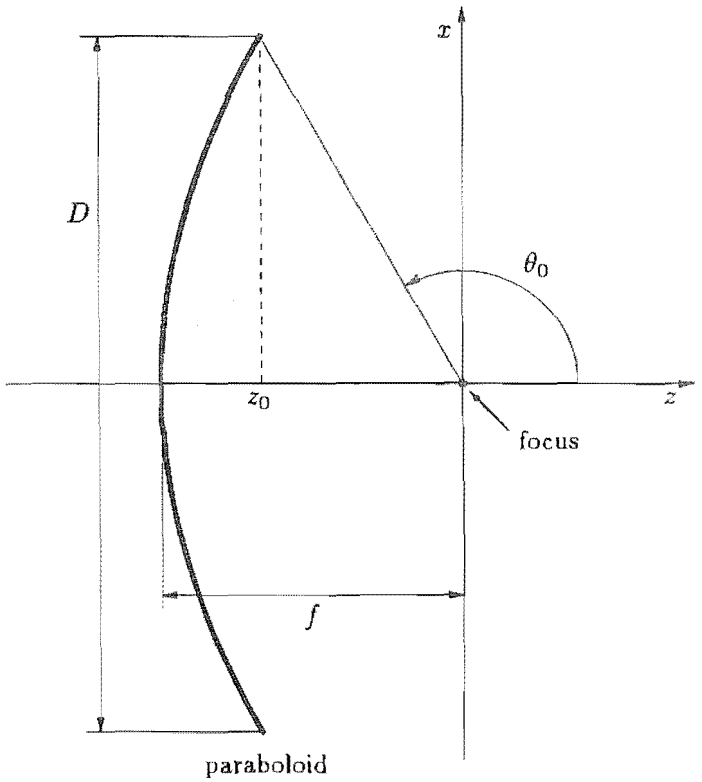
$$\begin{aligned} \hat{\mathbf{n}} &= \frac{-x\,\hat{\mathbf{x}} - y\,\hat{\mathbf{y}} + 2f\,\hat{\mathbf{z}}}{2[f(2f + z)]^{1/2}} \\ &= \cos\frac{\theta}{2}\cos\phi\,\hat{\mathbf{x}} + \cos\frac{\theta}{2}\sin\phi\,\hat{\mathbf{y}} + \sin\frac{\theta}{2}\,\hat{\mathbf{z}} \end{aligned} \tag{2.63}$$

Employing the laws of reflection, which are embodied in the first equation of (2.6), the ray is reflected in a direction parallel to

$$\hat{\mathbf{s}}_r = \hat{\mathbf{z}} \tag{2.64}$$

This demonstrates the important property of a paraboloidal reflector, which is that a set of rays diverging from the focus are transformed into a parallel set of rays. This

(a)



(b)

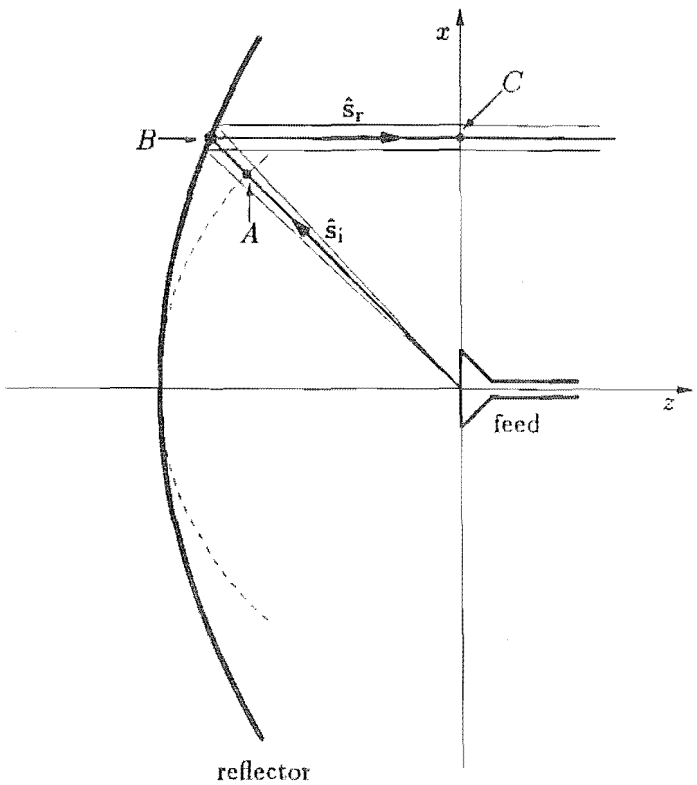


Figure 2.12 Cross-section through the axis of a paraboloidal reflector showing (a) its geometry and (b) a pencil of rays radiated by the feed.

property is independent of frequency, with the proviso that the frequency is high enough for the GO approximation to hold.

The far field pattern of the feed is here called the *feed pattern*. It is assumed that the feed pattern $\mathbf{E}_f(\theta;\phi)$ is defined over a sphere of radius f. The aperture field distribution is denoted by $\mathbf{E}_a(x,y)$. It can be equivalently expressed as a function of θ and ϕ, because there is a one-to-one mapping, from the angle at which a ray leaves the feed, to the point at which it intersects the aperture plane (after having been reflected). Because the reflected rays are parallel to the z axis, the mapping is given by substituting the second equation of (2.61) into (1.8):

$$\begin{aligned} \mathbf{E}_a(x,y) &= \mathbf{E}_a(\theta;\phi) \\ \text{where} \quad x &= \frac{2f\sin\theta\cos\phi}{1-\cos\theta} \\ y &= \frac{2f\sin\theta\sin\phi}{1-\cos\theta} \end{aligned} \tag{2.65}$$

Any vector component of the field at point A (on the sphere over which $\mathbf{E}_f$ is defined) is related, via (2.14), to a corresponding vector component of the field at point C (on the plane over which $\mathbf{E}_a$ is defined). The phase difference, between corresponding field vector components at points A and C, is proportional to the path length from A to C via B, where A, B and C lie on a single ray. From Figure 2.12, and employing (2.61) and (1.8), this path length is seen to be equal to

$$|z| + r - f = f \tag{2.66}$$

for any point B on the reflector. Therefore, there is a constant phase difference between corresponding vector components of $\mathbf{E}_f(\theta;\phi)$ and $\mathbf{E}_a(\theta;\phi)$ which is equal to kf, where k is the wave number (see (1.15)).

The polarization of the field along a ray is constant in free space, but usually changes at a point of reflection. For a paraboloid the GO rays are each reflected only once. Let the field, at point B, of the incident ray, be expressed by its spherical components $\mathbf{E}_i = (E_{i\theta}\,\hat{\boldsymbol{\theta}} + E_{i\phi}\,\hat{\boldsymbol{\phi}})$. Evaluating the second equation of (2.6) shows that the field, at point B, of the reflected ray is

$$\mathbf{E}_r = (E_{i\theta}\cos\phi + E_{i\phi}\sin\phi)\,\hat{\mathbf{x}} + (E_{i\theta}\sin\phi - E_{i\phi}\cos\phi)\,\hat{\mathbf{y}} \tag{2.67}$$

The amplitude of the field along a ray is given by the law of conservation of energy. The power flow of the feed's radiated field and the aperture field are related by

$$|\mathbf{E}_f(\theta;\phi)|^2 f^2 \sin\theta\, d\theta\, d\phi = |\mathbf{E}_a(x,y)|^2\, dx\, dy \tag{2.68}$$

where a pencil of rays, having a cross-sectional area of $f^2\sin\theta\,d\theta\,d\phi$ at point A, has an area of $dx\,dy$ at point C. The relationship between these two areas can be deduced from (2.65). When substituted into (2.68) it yields

$$|\mathbf{E}_a(\theta;\phi)| = \sin^2\left(\frac{\theta}{2}\right)|\mathbf{E}_f(\theta;\phi)| \tag{2.69}$$

The results of the previous paragraphs can now be combined to relate the aperture field distribution to the feed pattern. From (2.66), (2.67) and (2.69), the aperture field distribution can be expressed as [Rudge *et al.*, 1982, Sec. 4.3]

$$\begin{aligned} E_{ax}(\theta;\phi) &= \sin^2\left(\frac{\theta}{2}\right) e^{-jkf}\left(E_{f\theta}(\theta;\phi)\cos\phi + E_{f\phi}(\theta;\phi)\sin\phi\right) \\ E_{ay}(\theta;\phi) &= \sin^2\left(\frac{\theta}{2}\right) e^{-jkf}\left(E_{f\theta}(\theta;\phi)\sin\phi - E_{f\phi}(\theta;\phi)\cos\phi\right) \end{aligned} \tag{2.70}$$

where $\mathbf{E}_f(\theta;\phi) = [E_{f\theta}(\theta;\phi)\,\hat{\boldsymbol{\theta}} + E_{f\phi}(\theta;\phi)\,\hat{\boldsymbol{\phi}}]$ and $\mathbf{E}_a(\theta;\phi) = [E_{ax}(\theta;\phi)\,\hat{\mathbf{x}} + E_{ay}(\theta;\phi)\,\hat{\mathbf{y}}]$.

2.3.1.2 Feeds

The aperture field (2.70) is linearly polarized in the x direction when the feed pattern can be described by

$$\mathbf{E}_f(\theta;\phi) = E_0(\theta;\phi)(\cos\phi\,\hat{\boldsymbol{\theta}} + \sin\phi\,\hat{\boldsymbol{\phi}}) \tag{2.71}$$

where $E_0(\theta;\phi)$ is any complex scalar function. To maximize the peak gain of the antenna, the phase of the aperture field distribution must be uniform (see Sec. 2.2.5) and therefore, the phase of $E_0(\theta;\phi)$ must be independent of angle. Any symmetries in the feed pattern produce similar symmetries in the aperture field distribution.

If $E_0(\theta;\phi)$ is independent of ϕ, the feed is called a *balanced feed*[Thomas, 1976]. A feed for which the phase of $E_0(\theta;\phi)$ is constant is said to possess a *perfect phase centre*. An ideal example of a balanced feed, having a perfect phase centre, is a *Huygens source*, which consists of an electric and magnetic elemental dipole (Sec. 1.3.1) [Jones, 1954]. Practical balanced feeds for reflector antennas are the corrugated horn [Thomas, 1986] and other hybrid-mode feeds [Clarricoats and Poulton, 1977].

A horn feed and paraboloidal reflector can simultaneously illuminate the aperture with an x polarized field and an independent y polarized field. With such a feed, a circularly polarized aperture field can be generated by relating the linearly polarized fields in an appropriate way (see (1.19)).

A feed with a non-constant phase pattern phase$\{E_0(\theta;\phi)\}$ produces a non-uniform phase distribution in the aperture. This can be corrected by altering the shape of the reflector. For small phase deviations, the compensating correction to the second equation of (2.61) is [Cutler, 1947]

$$\delta r = \frac{1}{k}\frac{\text{phase}\{E_0(\theta;\phi)\}}{1-\cos\theta} \tag{2.72}$$

2.3.1.3 Practicalities

In practice, the reflector extends to a finite value of z, denoted by z_0 in Figure 2.12. The edge of the reflector forms the boundary of the aperture. An important property of a paraboloidal reflector is its f/D *ratio*, which is the ratio of the focal length to the diameter D of the aperture. From Figure 2.12, the half angle θ_0 subtended by the reflector aperture at the feed is given in terms of the f/D ratio by

$$\cot\theta_0 = \frac{1}{8(f/D)} - 2(f/D) \tag{2.73}$$

Paraboloidal reflectors with a large f/D ratio are called *shallow reflectors*.

In order to suspend the feed at the focus of the paraboloid, it must be supported by one or more struts. In a ray tracing analysis, both the feed and the struts intercept some of the rays before they reach the aperture. The feed and struts produce a shadow in the aperture plane, called aperture blockage (see Sec. 2.2.5), and also scatter the energy incident upon them. The combined effect is to reduce the gain of the main beam and increase the sidelobe levels.

As depicted in Figure 2.12(b), the feed is positioned so that its main beam is directed in the $\theta = 180°$ direction. Radiation from the feed at angles $\theta > \theta_0$ produce an aperture field distribution, in the unblocked parts of the aperture, given by (2.70). When employing the aperture field method of analysis, the aperture field distribution

is Fourier transformed to obtain an approximation to the far field pattern. However, radiation from the feed at angles $\theta < \theta_0$ constitutes spillover (Sec. 2.2.5), which also contributes to the antenna's radiation pattern. The main effect of the spillover is to increase the level of the sidelobes at angles close to the θ_0 direction. The effect, of both aperture blockage and spillover, on the copolar and cross polar radiation patterns may be different and can therefore cause depolarization in the far field.

2.3.2 Cassegrain antennas

A *Cassegrain antenna* consists of a feed, a subreflector and a main reflector. The prototype Cassegrain antenna has a paraboloidal main reflector, with a subreflector which is an axially symmetric section of a hyperboloid. As indicated in Figure 2.13, the subreflector is positioned so that one of its foci is coincident with the main reflector's focus, while the feed is positioned at the other focus of the subreflector. The aperture field distribution produced by a given feed in a prototype Cassegrain antenna is equal to the aperture field distribution produced by the same feed illuminating an *equivalent paraboloidal reflector*. The equivalent paraboloid has the same diameter as the main reflector and a focal length f_e of [Hannan, 1961]

$$f_e = f_m \frac{e+1}{e-1} \tag{2.74}$$

where f_m is the focal length of the main reflector and e is the eccentricity of the subreflector. Therefore, a balanced feed with a constant phase produces a uniformly phased, linearly polarized field on the aperture plane of a prototype Cassegrain antenna [Safak and Delogne, 1976].

In the design of a prototype Cassegrain antenna, a compromise must be reached between reducing spillover and maximizing the uniformity of the aperture field distribution. However, in shaped Cassegrain antennas, these are not necessarily opposing objectives [Rudge *et al.*, 1982, p. 248]. To minimize the spillover, a feed is selected with a large taper at the edge of the subreflector. The curvature of the subreflector is then chosen to distribute the feed's power evenly over the aperture. Finally the main reflector is then shaped to produce the desired phase distribution over the aperture.

In fact, to within the GO approximation, almost any circularly symmetric aperture amplitude and phase distribution can be produced by an arbitrary balanced feed and suitably shaped main reflector and subreflector [Galindo, 1964]. All such Cassegrain antennas have the property that rays emerging from a point feed are reflected, first from the subreflector and then from the main reflector, in a direction parallel to the axis of the antenna. Because of the symmetry, a Cassegrain antenna fed by a balanced feed has no cross polar field in the boresight direction [Rudge *et al.*, 1982, p. 248]. Diffraction analysis can also be employed in shaped reflector design [Wood, 1980, Sec. 7.2].

There is more aperture blockage in a Cassegrain antenna than in a paraboloidal antenna, because a subreflector is typically larger than a feed. However, only a short length of waveguide is required to connect the feed to the transmitting and receiving equipment, which can be conveniently positioned in the spacious volume behind the reflector. This results in minimal power loss in the waveguide. The position of the feed, close to the main reflector, also makes it easily accessible for adjustments.

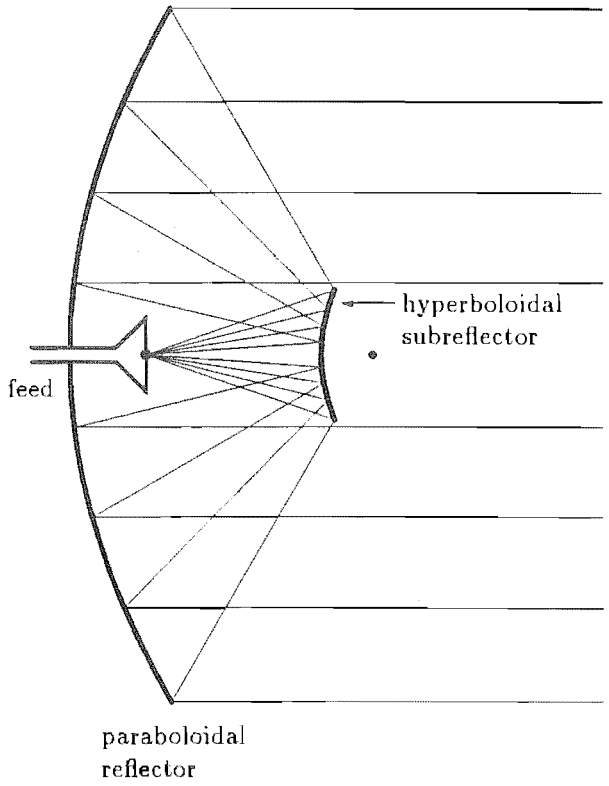


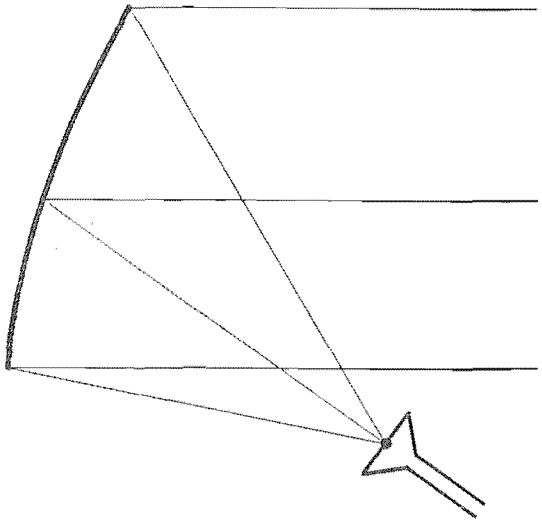
Figure 2.13 Geometry of a prototype Cassegrain antenna. The two dots indicate the positions of the foci of the hyperboloid and paraboloid.

2.3.3 Offset reflectors

A disadvantage common to both the paraboloid and the Cassegrain antenna, as described in the previous two sections, is the aperture blockage and scattering from the struts and either the feed or the subreflector. This blockage tends to reduce the main beam level and increase sidelobes (see Sec. 2.2.5). As is apparent from Figure 2.14, the blockage does not occur in *offset reflector antennas*.

A prototype single offset reflector antenna (Fig. 2.14(a)) consists of an asymmetrical section of a paraboloid, with a feed at its focus. The relationship between the feed pattern and the aperture field is described by (2.70). To avoid excessive spillover, the boresight of the feed is directed towards the centre of the reflector and is therefore no longer coincident with the z axis. This implies that a balanced feed (Sec. 2.3.1.2) produces an aperture field which is neither circularly symmetric nor linearly polarized in the x direction [Chu and Turrin, 1973]. However, for a given offset reflector, a feed can be designed to minimize the cross polarization component of the aperture field, using the matched feed concept [Rudge and Adatia, 1978]. As with paraboloidal reflectors, small phase errors in the feed pattern can be compensated for, by altering the shape of the reflector (see (2.72)).

(a)



(b)

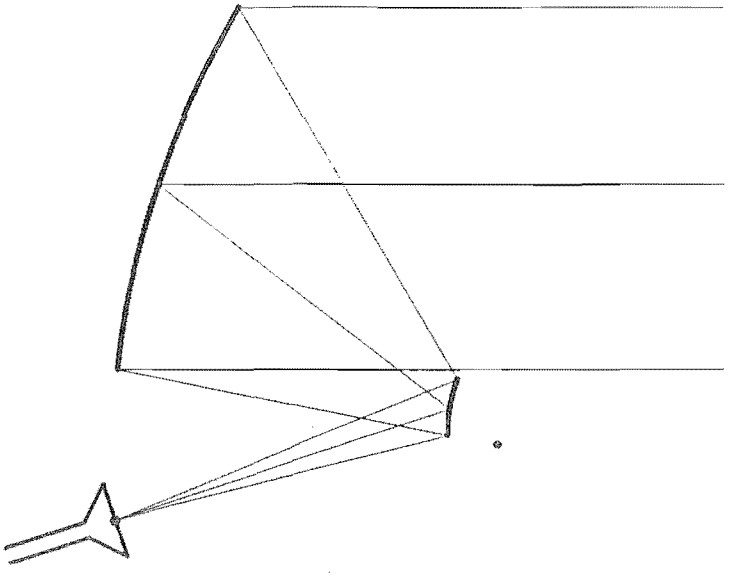


Figure 2.14 Geometry of offset reflector antennas: (a) single offset reflector configuration; (b) dual offset reflector configuration.

In a prototype dual offset reflector antenna (Fig. 2.14(b)), the main reflector and subreflector are asymmetrical sections of a paraboloid and a hyperboloid respectively. Like a Cassegrain antenna, one focus of the subreflector is coincident with the main reflector's focus, while the feed is centred on the other subreflector focus. Because an offset antenna is asymmetric, there is no requirement for the axes of the paraboloid and the hyperboloid to be coincident. When using a balanced feed, the depolarization effect of each reflector can be made to cancel by optimizing the angles between the axes of the paraboloid and hyperboloid [Fourikis, 1988].

As with Cassegrain antennas, dual offset reflectors can be shaped to produce any desired amplitude and phase distribution in the aperture, given any arbitrary feed [Galindo-Israel *et al.*, 1987]. Using the GO method, there are an infinity of solutions for the reflectors' shapes. The reflectors can alternatively be shaped to provide a uniform phase distribution over the aperture, with minimal cross polar field [Westcott and Brickell, 1979].

2.4 APPLICATIONS

Reflector antennas are employed in a wide range of situations, including terrestrial point to point communication, radar and communication with space craft. Two typical applications, in which high peak gain is a prime consideration, are the antennas used in radio astronomy (Sec. 2.4.1) and the earth station antennas used in satellite communications systems (Sec. 2.4.2). The discussion in the following sections summarizes the environmental and system factors which have an influence on the design of these antennas.

2.4.1 Radio astronomy

Radio astronomy is the study of radio emission from a variety of astronomical objects. The types of emission can be divided into three main groups [Christiansen and Högbom, 1985, Sec. 1.2]. Thermal emission is the result of black body radiation and produces a continuous spectrum of unpolarized fields. Synchrotron emission is radiated by highly relativistic electrons and is characterized by a pulsing, linearly polarized field, with a broad spectrum. Spectral line radiation is generated at specific frequencies by atomic and molecular processes. Information about the stars and other celestial objects can therefore be obtained by mapping the angular distribution of parameters such as brightness, polarization and frequency spectrum of the radio emission from the sky [Thompson *et al.*, 1986, Sec. 1.1]. A *radiotelescope* is an antenna, or an array of antennas, which is employed for the measurement of these parameters.

Most ground based observations have been made at frequencies lying between 30 MHz and 10 GHz. This range of frequencies constitute a *radio window*, within which the effect of earth's atmosphere on radio waves is comparatively minor. Radio waves with frequencies lower than about 30 MHz are reflected or severely refracted by the ionosphere (see Sec. 1.4.3). At frequencies above about 10 GHz, water vapour (including rain) in the air causes depolarization and power absorption of radio waves passing though the atmosphere. As the absorption increases, so to does the thermal noise generated by the water vapour (see Sec. 1.2.6). Within the window, most of the atmospheric effects on radio wave propagation discussed in Section 1.4 occur to some degree, depending on local conditions and on frequency. Other radio windows occur below about 300 kHz, and near to 90 GHZ and 280 GHz [Findlay, 1964].

A simple radiotelescope consists of a single high gain antenna. The region of sky to be observed is scanned by the main beam of the antenna, by moving either the whole antenna or just the feed. The angular resolution of the measurements is determined by the beamwidth of the main beam [Christiansen and Högbom, 1985, p. 9], which is in turn limited by the largest dimension of the aperture in terms of wavelengths (see (2.48)). To achieve an angular resolution comparable to that of even a small optical telescope, the dimensions of a single antenna would need to be impractically large.

However, an array of antennas has a beamwidth which is inversely proportional to the largest dimension of the array. The array's radiation pattern depends on both the individual antennas' radiation patterns and the geometrical arrangement of the antennas. The main beam can be steered electronically to scan an area of sky. An array of antennas can be simulated by suitably combining the data obtained from many different measurements, each made by a pair of antennas whose relative positions are altered between the measurements [Christiansen and Högbom, 1985, Sec. 1.7]. This arrangement is called a synthetic aperture radio telescope. It can only be employed when observing a source which does not vary significantly over the time taken to make all the measurements. It has the advantage of requiring only a small number of antennas.

Many sources of interest emit weak signals. The power measured by a radio telescope is proportional to its peak gain and the bandwidth of the receiving system. Interference can arise from other sources in the sky, thermal noise from the earth and radio waves from terrestrial systems operating within the same frequency band as the radiotelescope. Therefore, a radiotelescope is required to have a high peak gain and low sidelobes. It should also introduce a minimum amount of depolarization.

The paraboloidal reflector antenna is the most common antenna used for radiotelescopes [Christiansen and Högbom, 1985, p. 42]. The frequency at which the antenna operates can be changed by simply changing the feed and in this way a single reflector can be used over a wide range of frequencies. Although they can be shaped for optimum performance, Cassegrain antennas are not employed at low frequencies because, to operate effectively, the size of the subreflector is unacceptably large [Clarricoats and Poulton, 1977]. Shaping a single reflector can achieve only a uniform aperture phase distribution and is tailored for only one of the feeds. Therefore, the burden of achieving a desired aperture field amplitude and phase distribution, falls on the designers of the feeds.

2.4.2 Satellite communications systems

A satellite communications system consists of at least one satellite in orbit around the earth and at least two earth stations on the ground. The satellite receives signals from one or more transmitting earth stations, amplifies them and retransmits them at a different (usually lower) frequency to one or more receiving earth stations [Evans, 1986, Sec. 3]. In this way a single earth station can reliably reach another earth station which is nearly half way around the world.

Most satellite communications systems use *geostationary satellites* [Miya, 1981, Chap. 3]. A geostationary satellite is one which is in a 24 hour orbit in the equatorial plane and therefore its position appears (nearly) fixed to an observer on the ground. This makes it simple to track the satellite. Because the cost of launching a satellite depends on its weight, a satellite must be as light as possible. Electricity to run the satellite is generated from solar panels, which add to the weight, so the power requirements of the satellite must be kept to a minimum. The satellite must be reliable

enough to require no maintenance over its lifetime (more than 7 years [Gallois, 1987]) in the hostile environment of space.

2.4.2.1 Frequency reuse

A finite number of frequency bands in the radio spectrum have been assigned to satellite communications systems. The greatest use has been made of the bands at 4 and 6 GHz, as these suffer least from atmospheric influences [Evans, 1986, Sec. 3]. However, these bands have also been allocated to terrestrial users. To avoid interference from these users, many satellite systems are now operating with the bands near 11 and 14 GHz, even though heavy rain can cause severe attenuation at these higher frequencies. Whatever frequency bands are employed, there is a constant demand to utilize the available bands more effectively. This section discusses the main schemes for achieving *frequency reuse*, in which a given frequency is used simultaneously by many different signals.

An obvious way to reuse frequencies is to employ several geostationary satellites, each operating over the same frequency band. This can only be effective if the earth station antennas are directional enough to discriminate between signals from neighbouring satellites. Therefore, the minimum spacing between satellites is dependent upon the sidelobe levels of the earth station antennas [Miya, 1981, Sec. 3.2.1]. Presently, satellites operating in the 4 and 6 GHz bands are spaced at intervals of as little as 2° around the geostationary orbit.

Early satellites used global antenna beams which illuminated as much of the earth as possible. At a given frequency, each of these satellites can transmit or receive only one signal at a time. More recent satellites have achieved *spatial discrimination*, by employing several zone beams, each of which independently illuminates a different, non-overlapping region of the earth's surface. Therefore, at any given frequency, different signals can be simultaneously transmitted to, or received from, different regions. This achieves frequency reuse for a single satellite.

A satellite can achieve frequency reuse within each beam by employing *dual polarization*, in which one signal is propagated via either a linearly or a circularly polarized field, while simultaneously another signal is propagated via a field which has an orthogonal polarization. The level of interference between these signals depends on the depolarization effects of the atmosphere. It also depends on the depolarization characteristics of the antenna and feed systems on both the satellite and the ground. For frequencies below about 10 GHz, circular polarization is employed, because it is unaffected by Faraday rotation in the ionosphere. However, at higher frequencies, rain is the main cause of depolarization and linear polarization can be less affected than can circular polarization [Miya, 1981, Sec. 4.2.4].

2.4.2.2 Earth station antennas

The power transmitted by a satellite is kept to a minimum because of the limited power supply on the satellite, and also to minimize interference with terrestrial systems operating at the same frequency. Therefore it is important to achieve a high figure of merit (see Sec. 2.2.7) for an earth station antenna. It is also necessary to reduce the levels of the sidelobes as much as possible, because they directly influence the interference to and from other satellite and terrestrial systems.

In an effort to address the interference with neighbouring satellites, the CCIR has recommended a design objective for earth station antennas [CCIR, 1986a]. For an

antenna whose diameter exceeds 150 wavelengths (approximately 11 m at 4 GHz), the gain of 90% of the sidelobes must not exceed the envelope defined by

$$G(\theta) = 29 - 25\log\theta \quad \text{for } 1^\circ \leq \theta \leq 20^\circ \tag{2.75}$$

where G is in dB and θ is the angle from the boresight of the antenna. This requirement should be met for any direction which is within 3° of the geostationary orbit. In the USA, the FCC not only defines an envelope for directions toward the geostationary orbit, but also specifies that in all other directions the gain of the antenna shall fall below [Uyttendaele, 1986]

$$G(\theta) = \begin{cases} 32 - 25\log\theta & \text{for } 1^\circ \leq \theta \leq 48^\circ \\ -10 & \text{for } 48^\circ < \theta \leq 180^\circ \end{cases} \tag{2.76}$$

As part of the specification, a particular form of averaging may be employed to allow isolated high sidelobe peaks. Although the gain of the main beam is not specified by either (2.75) or (2.76), it is clear that the gain of the main beam must be as high as possible. For dual polarization systems to be effective, it is also required that their cross polar radiation patterns be minimized in directions corresponding to the locations of the satellites.

Cassegrain antennas are widely employed as large earth station antennas, because their reflectors can be shaped to optimize their radiation pattern. Producing an aperture field distribution which is tapered towards the edges of the aperture has the effect of lowering the sidelobe levels and decreasing the spillover past the main reflector (Sec. 2.2.5). However, the taper also reduces the gain of the main beam. The limiting factor in the reduction of side lobe levels of a Cassegrain antenna is the aperture blockage [CCIR, 1986c, Sec. 2]. Therefore, shaped offset dual reflector antennas have recently been employed, taking advantage of their unblocked apertures.

2.5 SUMMARY

Reflector antennas, operating at microwave frequencies, are able to produce highly directional radiation patterns. These antennas are therefore suited to applications which require a high peak gain, narrow main beam and low sidelobe levels.

Accurate methods of predicting antenna radiation patterns are essential to make possible the design of a candidate antenna. Such methods also help to provide an understanding of how the different components of the antenna contribute towards the radiation characteristics of the antenna as a whole. When applied to a reflector antenna, an analysis method must be provided with the radiation characteristics of the feed, the shape of the reflectors, and the relative positions and orientations of the reflectors and feed. An exact analysis can in principle be obtained by applying Maxwell's equations (1.9), but these are difficult to solve, even with the aid of a computer.

The simplest analysis technique is ray tracing (Sec. 2.1.1.4), in which the field is evaluated along independent rays emanating from the feed. The rays are straight, when in free space, and are reflected, according to the laws of reflection (Sec. 2.1.1.2), by reflectors. A high frequency approximation to the field along each ray is provided by the geometrical optics (GO) method (Sec. 2.1.1.3). The ray tracing method is simple in concept, but fails at caustics and in shadow regions. GTD (geometrical theory of diffraction; see Sec. 2.1.1.5) can be usefully invoked to supplement GO with extra rays

which account for diffraction. Realistic designs can be efficiently (from a computational point of view) implemented by combining GO and GTD.

In the current-integration method (Sec. 2.1.2), the field radiated by a reflector is calculated from the current distribution over the reflector's surface. Similarly, the field integration method (Sec. 2.1.3) predicts the field radiated by an aperture antenna, given the aperture field distribution. In both of these methods, the integrals can be simplified by applying approximations which hold in the far field region. In the case of the field integration method, the integral reduces to the Fourier transform operator (Sec. 2.1.3.2). The accuracy of both of these methods depends upon the accuracy with which the current or field distribution is known. The physical optics (PO) method (Sec. 2.1.2.3) is a current-integration method in which the reflector current distribution is that predicted by the GO method. Similarly, the aperture field method (Sec. 2.1.3.3) is a field integration method in which the aperture field distribution is taken to be that predicted by the GO method. Therefore, the aperture field method involves ray tracing and a Fourier transformation, both of which are easy to perform. This method is employed throughout the remaining chapters of this thesis. The PO method produces more accurate results, but it is computationally more demanding.

There are many other analysis techniques which have not been discussed in this chapter. The physical theory of diffraction (PTD) is an extension of PO, in much the same way that GTD is an extension of GO [Lee, 1977]. The GO method can be enhanced by employing equivalent edge currents to predict the field in GO shadow regions [James and Kerdemelidis, 1973]. Complex ray analysis (CRA) is a method of ray tracing through a complex coordinate space, where the rays are traced from the feed, to the reflectors and then into the far field region of an antenna [Hasselmann and Felsen, 1982]. The spherical wave expansion (SWE) method involves expressing the radiated field as a series expansion of vector spherical waves [Wood, 1980, Chap. 5]. The coefficients of these waves are determined by knowledge of a current or field distribution over a closed surface. In the spherical near field GTD method, GTD is employed to predict the field on a sphere which just encloses the antenna. This field is then transformed into the far field, by employing the SWE method [Bach and Viskum, 1987]. For a given antenna, the different analysis methods can complement each other. For example, an analysis by James [1980] employs the aperture field method to predict the radiation pattern at angles close to the main beam, and the GTD method is invoked to predict the remainder of the radiation pattern.

The main configurations of high gain reflector antennas are discussed in Section 2.3. From a GO analysis, they all have the common property that most of the radiation transmitted by the feed emerges from the antenna aperture as a collimated beam. A more rigorous analysis, that considers diffraction effects, reveals that an antenna radiates power in all directions. The radiation pattern exhibits a main beam, containing most of the radiated power, and several sidelobes. The width of the main beam is inversely proportional to the diameter of the aperture in wavelengths.

Many factors influence the peak gain and the sidelobe levels of the radiation pattern of a reflector antenna (Sec. 2.2). The peak gain of an antenna can be maximized by having a uniform amplitude and phase distribution in the aperture, minimizing the blockage in the aperture and minimizing the spillover. The level of the sidelobes can be decreased by having an aperture field amplitude distribution which tapers towards the edge of the aperture, thereby minimizing spillover, and by minimizing scattering from objects which block the aperture. For applications requiring high polarization purity, the distribution of the cross polar component of the aperture field should be kept to a

minimum.

The paraboloidal and Cassegrain reflector antennas both employ axially symmetric reflectors. When fed by balanced feeds, they both suffer minimally from depolarization. However, they do suffer from aperture blockage by the struts and the feed or subreflector. The reflectors of the Cassegrain antenna can be shaped to minimize spillover and to provide any desired aperture field distribution. Offset reflector antennas have the advantage of possessing an unblocked aperture, but cause significant depolarization when fed by a balanced feed. Shaping the reflectors and employing different feeds can help to alleviate this difficulty.

Two important applications for high gain reflector antennas are radio telescopes (Sec. 2.4.1) and satellite communications earth stations (Sec. 2.4.2). In both of these applications the antenna is required to receive weak signals from a single direction. Therefore, a high gain is required in this direction, to maximize the strength of the received signal. Low sidelobe levels are also required, so that the strength of interfering signals from all other directions is minimized. When these two requirements are met, the antenna has a high figure of merit (Sec. 2.2.7). Each application requires a narrow main beam, to restrict the field of view of the antenna. When an earth station antenna is transmitting, it should exhibit these same properties, but for different reasons: a high peak gain is required to maximize the power radiated in the desired direction and low sidelobe levels are required to minimize interference with other systems. In both applications, when either transmitting or receiving, it is important to realize high polarization purity in the main beam. The characteristics of the nulls and phase of the radiation pattern are of secondary importance, except for antennas designed to reject a strong source of interference from a specified direction.

CHAPTER 3

RETRIEVAL OF APERTURE FIELD PHASE

The radiation pattern produced by a given reflector antenna is determined by the following factors:

1. The radiation characteristics of the feed.
2. The geometry of the antenna, including the shape (profile) of the reflectors and their position relative to the feed.
3. The environment in which the antenna is to operate, including the propagation characteristics of the atmosphere (Sec. 1.4) and objects (e.g. the ground, nearby structures) which scatter radiation.

The specification for a radiation pattern incorporates safety factors allowing for the anticipated environmental effects, which are usually out of the designer's control. Therefore, if the radiation pattern fails to meet its design specifications, the implication is that the antenna itself must differ significantly from its design. The radiation characteristics of the feed can be tested in one of the standard types of antenna measurement range [IEEE, 1979]. However, because the antenna is usually too large to be moved, its geometry must be measured on site. One of the characteristics of an antenna which is straightforwardly measurable is its amplitude pattern (Sec. 1.2.2). The purpose of this chapter is to show that, in principle, information about the antenna's geometry can be inferred from a measured amplitude pattern of the antenna. Two steps are involved. Firstly, the copolar aperture field distribution is retrieved (estimated) from the measured data, and secondly, the defects of the antenna geometry are deduced from the phase of the copolar aperture field distribution.

Some of the notation utilized in this chapter is introduced in Section 3.1. This section also outlines the likely causes of geometrical defects of high gain reflector antennas and indicates how they can be corrected. Section 3.2 shows how the geometrical defects can be deduced from the copolar aperture phase distribution. Methods for inferring the antenna geometry directly, and for measuring the radiation from the antenna, are presented in Section 3.3. Estimation of the copolar aperture phase distribution from the copolar amplitude pattern requires the solution of the Fourier phase problem, which is discussed in Section 3.4. Previously reported methods for recovering the copolar aperture phase distribution from a measured copolar amplitude pattern are reviewed in Section 3.5.

It is convenient to note here the meanings of the word *holography* because, although two of the methods described in this chapter are described as holographic, they are quite different from each other. Gabor [1949] coined the word 'hologram' to mean a photograph of the optical diffraction pattern caused by interference between a 'background wave' and a 'secondary wave'. The reason for introducing the term was that a hologram

is an intensity (or amplitude) recording, while a diffraction pattern is complex valued. Gabor [1949] shows how (the amplitude and phase of) the secondary wave can be reconstructed from the hologram. Papi *et al.* [1971] state that 'microwave holography' is an extension of optical holography to the microwave field. Napier and Bates [1973] and Bennett *et al.* [1976] have instituted such a microwave holographic approach to determine the aperture field distribution of an antenna from a record of only the amplitude of the interference between the radiation patterns of a reference antenna and the antenna under test. However, as pointed out by Morris [1985], people have retained the name when describing more recently developed methods in which both amplitude and phase are recorded. Therefore Anderson [1977] and Tricoles and Farhat [1977] define microwave holography as reconstructing a field either from records of its amplitude and phase or from a record of only its amplitude. This definition embraces all of the methods of field measurement discussed in Sections 3.3 and 3.5. Note that reconstruction of fields from records of their amplitudes and phases requires quite different techniques than does reconstruction of fields from records of only their amplitudes. In this thesis I call the original microwave holography 'amplitude holography'. Reconstructing the aperture field distribution, from a record of both the amplitude and the phase of the radiation pattern of the test antenna, I classify as 'complex holography'.

3.1 GEOMETRICAL DEFECTS OF REFLECTOR ANTENNAS

Before discussing the various kinds of geometrical defects encountered in high gain reflector antennas, the following scenario is presented to introduce terms which are used in this section and throughout the remainder of this thesis:

1. An antenna is designed to produce a particular far field pattern. This is achieved by utilizing a given feed and one or more reflectors positioned and shaped according to the *design geometry.* The resultant aperture and far fields are referred to as the *design fields.* The word 'design' is here equivalent to 'desired', 'optimum', or 'ideal'.

2. The antenna is constructed to reproduce as closely as possible the design geometry. The result is an *actual geometry*, which produces *actual fields* in the aperture and the far field region.

3. Differences between the design and actual geometry are here called *geometrical defects.* They include *shape defects* of the main reflector and subreflector, and *displacements* of the relative positions of the main reflector, the subreflector and the feed. The resulting differences between the actual and design fields are called *field deviations.* The term *aperture phase deviations* refers to differences in the phase between the copolar components of the actual and design aperture fields. An estimate of the geometrical defects can be inferred from either direct measurement of the actual geometry (Sec. 3.3.1), or an estimate of the aperture phase deviations (Sec. 3.2). The word 'error' is not used here, because it is reserved for parameters describing convergence properties of an algorithm (e.g. (3.60)).

4. If measurement reveals that the actual far field pattern does not meet its specifications, the antenna must be adjusted in some way to produce a far field which is closer to the design. In particular, any geometrical defects which contribute to significant field deviations, should be corrected. This might involve reshaping

a reflector, or realigning the reflectors and feed of the antenna. The result is referred to as the *corrected geometry*, and it produces *corrected fields* in the aperture and the far field region. The correction is deemed successful if a subsequent measurement of the antenna's far field radiation pattern reveals that it now does meet its specifications.

There are many different mechanisms which give rise to geometrical defects of high gain reflector antennas. Some of the defects can be minimized by appropriate mechanical design of the antenna, while others can be corrected after the antenna has been constructed. The various deforming mechanisms are summarized in the following paragraphs.

The main reflector surface of a high gain antenna typically consists of many reflecting panels, each of manageable size. The panels are supported by a frame which provides structural stiffness. In practice, the overall shape of the reflector can never be manufactured to correspond exactly to the design shape. However given sufficient resources, the reflector shape can be manufactured to any required tolerance. A rule of thumb figure for the permissible tolerance of the main reflector is 1/32 to 1/16 wavelengths [Blake, 1984, p. 281]. For a 30 m diameter antenna, operating at a maximum of 6 GHz, this corresponds to an accuracy of about 2 mm, or one part in 15 000. The fabrication of larger antennas, and ones which operate at higher frequencies, involves correspondingly higher accuracies.

In many antennas, the position of the main reflector panels, the subreflector and the feed can be adjusted after the antenna has been constructed [e.g. Godwin *et al.*, 1986]. Therefore, any misalignments of these components during assembly can in principle be corrected later. Individual panels which are excessively deformed can sometimes be replaced or reshaped. Alternatively, the subreflector can be reshaped to compensate for the deformations in the main reflector [Milner and Bates, 1980]. For antennas fed by an array, the reflector deformations can be compensated for by appropriately altering the amplitude and phase of the signal fed to each element of the array [Rudge and Davies, 1970; Cornwell and Napier, 1988].

Not only do the individual panels have to be accurately shaped, but the frame must be rigid enough to keep the deflections due to gravitational and wind loading of the main reflector to within the required tolerance. The gravitational deflections, due to the weight of the antenna, change in a predictable way as the antenna is directed to different elevations. In a paraboloidal reflector the gravitational deflections are *homologous* to first order and *astigmatic* to second order [von Hoerner and Wong, 1975]. The antenna remains paraboloidal for homologous deflections, but its focal length varies as the antenna points to different elevations. When suffering from astigmatic deflections, the cross-sections, through different diameters of the actual reflector, are parabolas with different focal lengths. In what is known as homologous design of reflectors [Findlay, 1971], the astigmatic and higher order deflections are minimized. The position of the subreflector or feed must then be appropriately adjusted with elevation.

The forces due to wind vary in both magnitude and direction with time. The resulting deflections can be reduced by utilizing a more rigid antenna structure, or by employing perforated panels which reduce the wind resistance. Thermal gradients across the reflector surface cause buckling of the panels. For smaller antennas, both of these sources of variable geometrical defect can be reduced by enclosing the antenna in a radome. However, for larger antennas, these effects must be minimized by suitable choice of materials and mechanical design.

3.2 RELATING GEOMETRICAL DEFECTS TO APERTURE PHASE DEVIATIONS

The following sections demonstrate how geometrical defects of a reflector antenna can be inferred from the aperture phase deviations. The implication is that field measurements can provide an alternative to direct measurement of an antenna's geometry for assessing how serious are the antenna's geometrical defects.

Sections 3.2.1 and 3.2.2 describe the relationship between geometrical defects and the resulting aperture phase deviations, for a paraboloidal reflector. More complicated antenna types are discussed in Section 3.2.3. The only defects considered are those which are small compared to the overall dimensions of the antenna. This is appropriate because larger defects could be readily identified with the aid of a template, if not from visual inspection of the antenna reflector surfaces. The geometrical optics approach which is employed in the following sections is consistent with the standard aperture field method of antenna analysis (see Sec. 2.1.3.3).

3.2.1 Reflector shape defects

Figure 3.1 shows a cross-section through a portion of a paraboloidal reflector suffering from a shape defect. The copolar component of the aperture field distribution is taken to be defined by any arbitrary but constant polarization unit vector lying in the aperture plane (cf. (2.56)). The copolarization unit vector is usually conveniently chosen to coincide with the polarization designed to be radiated by the antenna. Consider the ray which passes through the point C_d, at position (x, y), in the aperture plane. The design ray $\overline{AB_dC_d}$ and reflector profile (both drawn as dashed curves) are as described and analysed in Section 2.3.1.1.

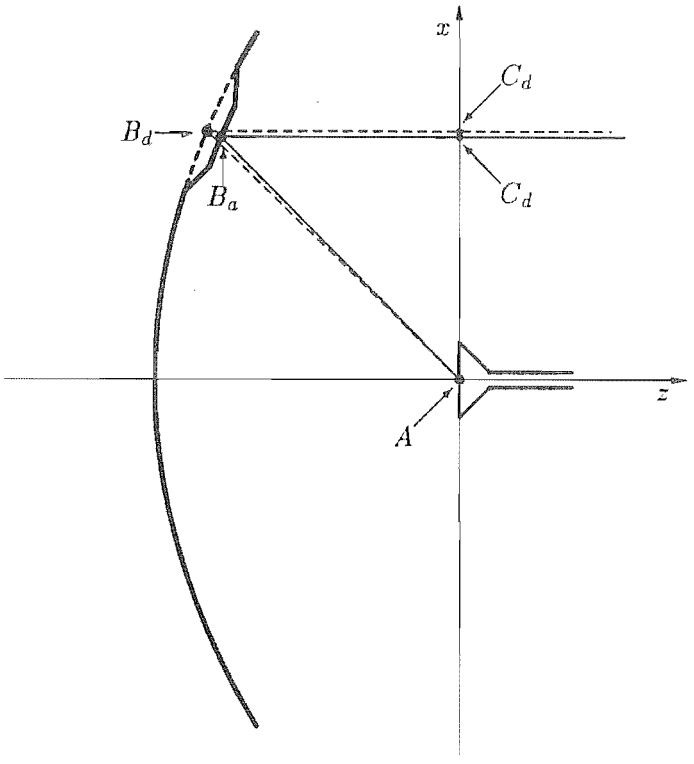
The point B_d on the design reflector is displaced by $\Delta n(x, y)$ in the direction of the normal to the design surface. The resulting point B_a lies on the actual reflector surface (solid curve). The actual ray through B_a is $\overline{AB_aC_a}$.

It is assumed that $\Delta n << f$, where f is the focal length of the paraboloid. It is also assumed that the design and actual reflector surface normals, at B_d and B_a respectively, are approximately parallel. Under these conditions, the rays $\overline{AB_dC_d}$ and $\overline{AB_aC_a}$ are approximately coincident, as indicated in Figure 3.1(a). Similarly, a pencil of rays surrounding $\overline{AB_dC_d}$ is approximately coincident with the corresponding pencil of rays surrounding $\overline{AB_aC_a}$. These two pencils of rays leave the feed over approximately the same solid angle and therefore transport approximately the same power. When they intersect the aperture plane, their cross-sectional areas are approximately equal. Therefore, according to the rules of the ray tracing method (Sec. 2.1.1.4), the amplitudes of the copolar fields are approximately equal at the almost coincident points C_d and C_a. Applying the same reasoning to all pencils emitted by the feed shows that the amplitudes of the copolar actual and design aperture fields are approximately equal over the whole aperture. However, the path lengths of the design and actual rays can be significantly different. Inspection of Figure 3.1(b) reveals that the path length difference Δl, for a ray leaving the feed at an angle θ from the z axis, is [Slater, 1970]

$$\Delta l = -2\,\Delta n(x, y) \cos\left(\frac{180^\circ - \theta}{2}\right) = -2\,\Delta n(x, y) \sin\left(\frac{\theta}{2}\right) \tag{3.1}$$

which introduces an aperture phase deviation $\Delta\psi$, which is defined in Section 3.1. The shape of a paraboloid, specified by (2.61), ensures that the distribution of $\Delta\psi$ over the

(a)



(b)

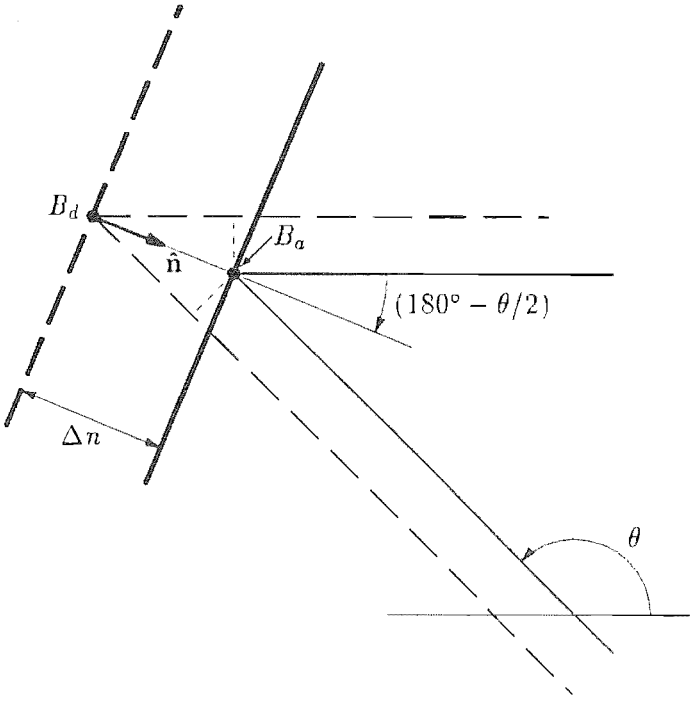


Figure 3.1 Cross-section through a paraboloidal reflector having a shape defect: (a) rays from the feed to the aperture plane; (b) enlargement of the region near to the defect. The actual reflector and ray (solid curves) and the design reflector and ray (dashed curves) are shown for each case.

aperture plane is given by either [Ruze, 1966]

$$\begin{aligned} \Delta\psi(\theta;\phi) &= 2k\,\Delta n(\theta;\phi)\sin\left(\frac{\theta}{2}\right) \\ \text{or}\quad \Delta\psi(x,y) &= \frac{4kf\,\Delta n(x,y)}{(4f^2+x^2+y^2)^{1/2}} \end{aligned} \tag{3.2}$$

using the notation established in Section 2.3.1.1.

For a shallow paraboloid, $\sin(\theta/2) \approx 1$ and

$$\Delta\psi(x,y) = 2k\,\Delta n(x,y) \tag{3.3}$$

Parini *et al.* [1989] have reported an alternative method of determining the geometrical defects of a paraboloidal antenna when the phase of the copolar aperture field distribution is available. Like the method described above, it uses a GO approach to trace actual rays from the feed to the aperture plane. Where it departs from the above method is that the actual rays are not assumed to approximately coincide with the design rays. This implies that it can be employed when the reflector has relatively large defects and when the feed is not at the reflector's focus. Consider the ray $\overline{AB_aC_a}$. From standard geometry, the point B_a can be uniquely determined from the direction of $\overline{B_aC_a}$ and the total length of $\overline{AB_aC_a}$, where the positions of A and C_a are known. The direction of $\overline{B_aC_a}$ is given by the gradient of the copolar field phase distribution in the aperture plane at C_a because the wavefronts, which are surfaces of constant phase, are locally perpendicular to the rays (Sec. 2.1.1.3). The difference in length between two neighbouring rays is given by k times the phase difference between the copolar fields at the points at which they intersect the aperture plane. By iteratively comparing all rays with their neighbours, their lengths can all be given in terms of the length of any one ray, which must be measured directly. For every ray which is considered in this way, the position of a point B_a can be calculated, so for many rays, the shape of the actual reflector can be determined. This shape can then be compared with the design shape to determine any shape defects.

3.2.2 Feed displacement

Figure 3.2 shows a cross-section through a paraboloidal reflector which is fed by a misaligned feed. The phase centre of the feed is displaced from the reflector focus by Δx, Δy and Δz, in the x, y and z directions respectively. Provided that the displacement is small compared to the focal length of the reflector, the design and actual rays are approximately parallel. Therefore, by appealing to the rationale presented in the previous section, the amplitude deviation of the copolar aperture field can be neglected.

The direction of the ray $\overline{A_aB}$, as it leaves the feed, is conveniently described in terms of the spherical coordinate angles θ and ϕ. From the geometry of Figure 3.2, where $\overline{A_aB}$ and $\overline{A_dB}$ are assumed parallel, the path length difference between the design and actual rays is

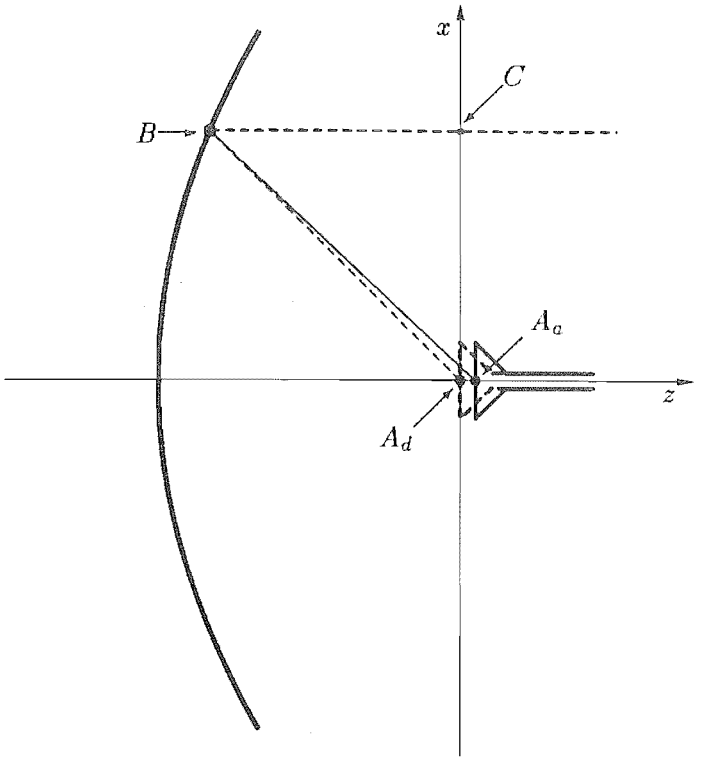
$$\Delta l = -\Delta x\sin\theta\cos\phi - \Delta y\sin\theta\sin\phi - \Delta z\cos\theta \tag{3.4}$$

The resulting aperture phase deviation is given by [Ruze, 1965]

$$\Delta\psi(\theta;\phi) = k(\Delta x\sin\theta\cos\phi + \Delta y\sin\theta\sin\phi + \Delta z\cos\theta) \tag{3.5}$$

An antenna with an axially displaced feed (i.e. $\Delta z \neq 0$) is called a *defocused antenna.* When this is the only geometrical defect in the antenna, the distribution of the aperture

(a)



(b)

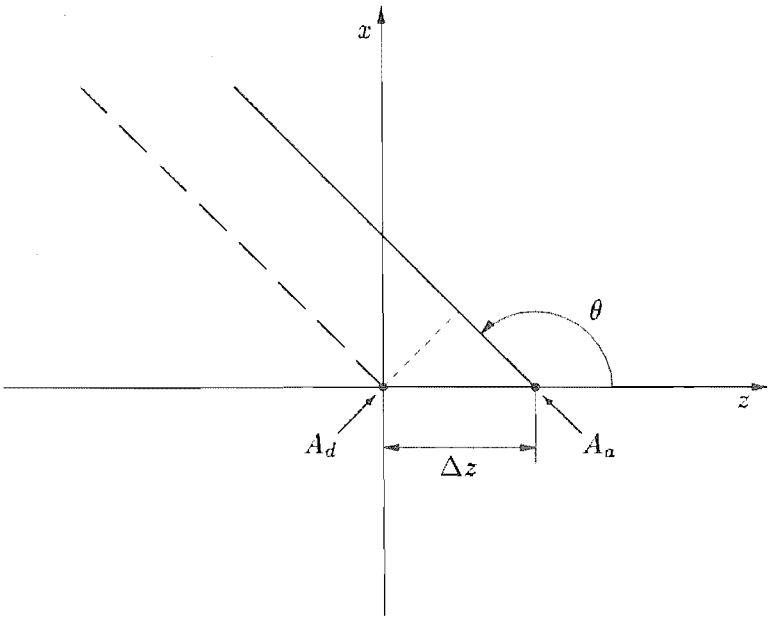


Figure 3.2 Cross-section through a paraboloidal reflector with a displaced feed. Both an actual (solid line) and a design (dashed line) ray are shown (a) from the feed to the aperture plane and (b) in an enlargement of the region near to the feed.

phase deviation is circularly symmetric. It is therefore convenient to define a normalized radius ρ in the aperture by

$$\rho = \frac{(x^2 + y^2)^{1/2}}{D/2} \tag{3.6}$$

where D is the diameter of the aperture. Upon substituting (2.61) and (1.8), into (3.5), the distribution over the aperture plane of the aperture phase deviation for a defocused antenna, for which $\Delta x = \Delta y = 0$, is given by [Chu, 1971]

$$\Delta\psi(\rho) = k\,\Delta z \left[\frac{2\rho^2}{\rho^2 + (4f/D)^2} - 1\right] \tag{3.7}$$

This equation can be approximated by a circularly symmetric quadratic function which is equal to the expression in (3.7) at both the centre and the edge of the aperture:

$$\Delta\psi(\rho) = k\,\Delta z\,\frac{2\rho^2}{1 + (4f/D)^2} - k\,\Delta z \tag{3.8}$$

The maximum difference between the expressions in (3.7) and (3.8), for a reflector whose f/D ratio is 0.33, is around 12% of the difference between the maximum and minimum values of $\Delta\psi$ [Johnson *et al.*, 1973, Fig. 29]. The two expressions agree more closely for shallow reflectors.

A non-constant distribution of phase deviation over the aperture implies that the copolar actual and design radiation patterns differ in both amplitude and phase. However, a phase deviation which is constant over the aperture implies that the copolar actual and design radiation patterns differ by only a constant phase. Because the absolute phase of an electromagnetic field has no operational meaning, constant aperture phase deviations, such as the term $k\,\Delta z$ occurring in (3.8), can always be ignored.

3.2.3 Inferring geometrical defects from aperture phase deviations

The scenario introduced in Section 3.1, implies that it is mandatory to estimate the geometrical defects, so that they can be corrected. Reflector shape defects and feed displacements in a paraboloidal antenna are related to the resulting aperture phase deviations in Sections 3.2.1 and 3.2.2 respectively. Therefore, if an estimate of how the phase deviation is distributed throughout the aperture is available, the geometrical defects can themselves be estimated. For example, the amount and direction of feed displacement from the focus can be determined by finding the values of Δx, Δy and Δz which allow the best fit of (3.5) to any available aperture phase deviation data.

The ray tracing approach developed above, in Sections 3.2.1 and 3.2.2, can also be employed for other types of reflector antenna, to determine a relationship between the geometrical defects and the aperture phase deviations. Application of this analysis to a prototype Cassegrain antenna reveals that an axial displacement of the feed produces an aperture phase deviation of the same form as (3.5). An axial displacement of the subreflector produces an aperture phase deviation consisting of two terms: one due to the path length differences between the feed and the subreflector; the other due to the path length differences between the subreflector and the main reflector. In a typical Cassegrain antenna, an axial displacement of the subreflector causes about 16 times the aperture phase deviation due to an equal axial displacement of the feed [Rudge *et al.*, 1982, p. 169]. The analysis can also be applied to shaped Cassegrain antennas to relate a feed or subreflector displacement to a distribution of phase deviation in the aperture [Claydon, 1970].

In general, a mathematical relationship between the geometrical defects and the aperture phase deviations need not be explicitly calculated. The geometrical defect giving rise to a phase deviation at a given point in the aperture can usually be estimated to an acceptable accuracy by tracing a ray backwards along trajectory of the design ray from the given point to the feed. The phase deviation is caused by one or more of the following:

1. A path length difference due to a shape defect of main reflector at point at which the ray reflects from it.

2. A path length difference due to a shape defect of the subreflector, if one is present, at the point at which the ray is reflected from it.

3. A path length difference due to a displacement in the position of the feed.

4. A phase deviation in the radiation characteristics of the actual feed, at the angle at which the design ray leaves the feed.

This ambiguity implies that it is not always possible to determine the geometrical defects. It does suggest, however, that correction of an aperture phase deviation can be achieved in a variety of ways. For example, a shape defect in the main reflector can be corrected directly, or, as can sometimes be more convenient, compensated by introducing appropriate shape defects into the subreflector [von Hoerner, 1976; Milner and Bates, 1980; Godwin *et al.*, 1985]. An advantage of knowing the aperture phase deviations is that displacements of the feed or subreflector can be estimated, as well as shape defects of the reflectors. Furthermore, phase deviations in the radiation characteristics of the feed can also be detected. These can then be compensated by altering the shape of the reflectors, as intimated in Section 2.3.1.2.

By definition (Sec. 3.1), the distribution of aperture phase deviation can be calculated by calculating the difference between the design and the actual copolar aperture field phase distributions. However, in most cases, the design copolar aperture field has a uniform phase distribution (for reasons discussed in Sec. 2.2.5). This implies that the distribution of the aperture phase deviation is equal to the actual copolar aperture field phase distribution. Therefore, the theory developed in this section and Sections 3.2.1 and 3.2.2 can usually be utilized to determine the geometrical defects from knowledge of the actual copolar aperture field phase distribution alone.

A limitation to the above theory, is that the geometrical defects, and the implied corrections, must be small enough not to significantly alter the directions of the rays from the feed to the aperture. Application of the theory is also limited by the approximations inherent in the ray tracing method, which does not account for diffraction between the feed and the aperture plane. However, the accuracy of the resulting estimate of the geometrical defects can always be assessed by comparing the aperture field deviations (inferred from measurements of the field radiated by the antenna: see Secs. 3.3, 3.5 and Chap. 4) with the aperture field deviations calculated by a diffraction analysis incorporating the estimated geometrical defects.

3.3 MEASUREMENT METHODS

As intimated in Section 2.4, an important measure of an antenna's performance is its radiation pattern. Geometrical defects of the antenna usually degrade the radiation pattern. When they are significant enough to cause the radiation pattern to fail its

specifications, the defects must be corrected, by employing any of the methods outlined in Section 3.1. This section reviews established measurement and computational procedures for estimating radiation patterns and geometrical defects of high gain reflector antennas. These procedures involve direct measurement of either the geometry of, or the field radiated by, any such antenna. Although some components of the antenna can be checked at the factory, a complete final check must be performed on site, after the antenna is installed, in case certain components have been displaced.

At first sight, the most obvious method of providing an estimate of the geometrical defects of an antenna is to measure the actual geometry and compare it with the design geometry. The position and orientation of the reflectors and feed can be determined by measuring the three-dimensional position of at least three points on each of these components. Determining the shape of a reflector requires the accurate measurement of the three-dimensional position of hundreds, or even thousands, of points on the reflector. The shapes of small reflectors, including subreflectors, can be measured by reference to a template [Findlay, 1971]. Different methods for directly measuring the shape of large main reflectors are summarized in Section 3.3.1. Some of these measurement methods can be adapted to measurement of the shapes of subreflectors, and to determination of the relative positions of reflectors and feeds.

An estimate of the geometrical defects of an antenna can alternatively be determined from the aperture field distribution by employing the method described in Section 3.2. The aperture field can be measured directly by using the planar near field scanning technique which, along with other near field scanning techniques, is outlined in Section 3.3.2.

The radiation pattern of an antenna can be directly determined by measuring either the Fourier Fresnel or far field. Techniques for making these measurements are outlined in Section 3.3.3. The results of these measurements can be processed to provide an estimate of the copolar aperture field distribution which can in turn provide an estimate of its geometrical defects. In the complex holography technique (Sec. 3.3.3.2), both phase and amplitude measurements are made of the copolar radiation pattern, so that the copolar aperture field distribution can be straightforwardly computed by inverse Fourier transformation. The copolar aperture field distribution can also be inferred merely from measurements of the amplitude of the copolar radiation pattern (Sec. 3.3.3.1), by employing the phase retrieval techniques described in Section 3.5 and Chapter 4.

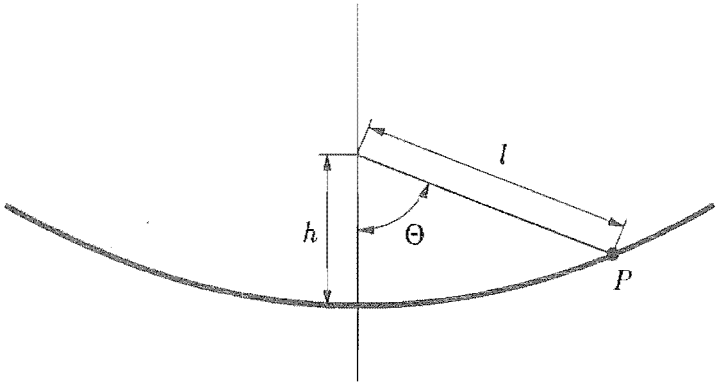
In Section 3.3.4 the different measurement techniques are compared. It is argued that in many situations it is preferable to infer the geometrical defects of an antenna from measurements of only the copolar amplitude pattern of the antenna.

3.3.1 Measurement of reflector shapes

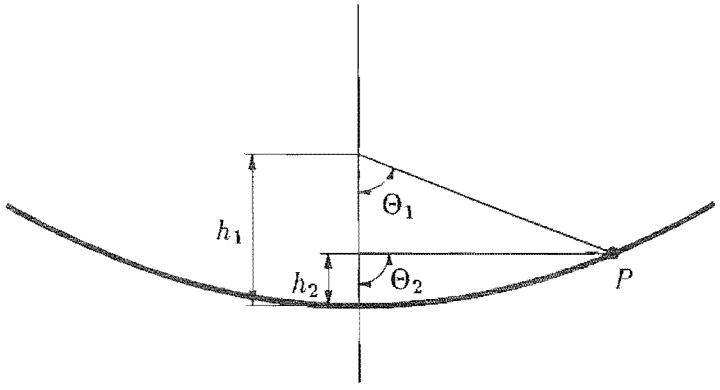
Direct measurement of a large main reflector can be achieved by mechanical and/or optical means. Any suitable measurement method must be able to determine the three-dimensional position of any point on the reflector surface to within the desired accuracy. Because the positions of a large number of points are required, it is preferable for the method to be incorporated into an automated system [IEEE, 1979, p. 64].

Figure 3.3(a) illustrates the use of a survey tape and theodolite to determine the position of point P on the surface of the reflector. The theodolite, positioned a known height h above the centre of the reflector, measures the angle Θ between the antenna axis and the line connecting the theodolite and P. The tape measures the distance l

(a)



(b)



(c)

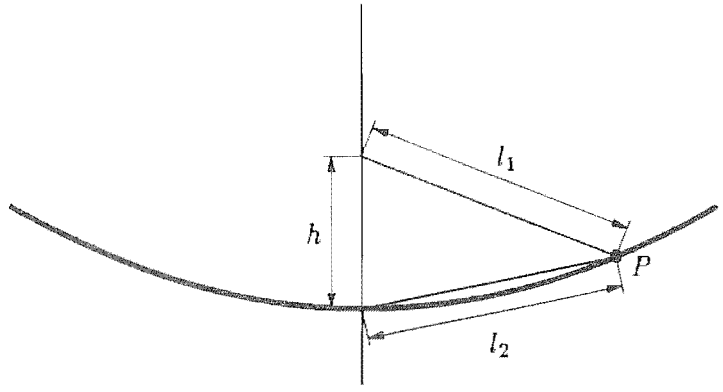


Figure 3.3 Various methods for measuring the shape of a main reflector. The indicated lengths and/or angles are measured or determined by the following instruments: (a) tape and theodolite; (b) pentaprisms; (c) modulated laser beam.

to P from the theodolite. This method can have an rms accuracy of 1 mm when l is as large as 20 m [Findlay, 1971].

An alternative method employs two pentaprisms which can each be moved along the axis of the antenna (Fig. 3.3(b)). A pentaprism deflects an incident light ray through a fixed angle Θ. The value of Θ is relatively insensitive to the orientation of the prism [Kelly *et al.*, 1970]. The height of each pentaprism is altered until it produces a centred image of a target placed at P, when viewing along the axis of the antenna. The height h and angle Θ of each pentaprism can then be used to locate the target. Slater [1971] reports employing this method on a 25 m diameter antenna, obtaining an rms repeatability of 0.51 mm for measurements of points on the reflector.

An alternative to a tape for length measurements is a modulated laser beam, which is directed by a mirror to a target on the reflector surface. The target, which is a small optical corner cube, reflects the beam back to the mirror. The difference between the phases of the modulation on the reflected and incident rays gives a measure of the distance from the mirror to the target. Utilizing this technique, Payne [1973] was able to achieve an accuracy of 0.076 mm over a length of 60 m. Figure 3.3.1(c) illustrates how two of these mirrors can be positioned at two fixed points to determine the position of the target. For a paraboloidal reflector, such measurements can reveal the position of its focus relative to the two fixed points [Anderson and Groth, 1963], thereby facilitating the positioning of the feed or subreflector.

Payne *et al.* [1976] describe a technique which consists of measuring the curvature of the reflector as a function of distance along a path on the reflector surface. The profile of the reflector along that path can be calculated by integrating this function twice. An accurate measure of curvature is provided by a three wheeled cart with an electronic depth gauge mounted at the midpoint of the cart. The cart is rolled from the centre of the reflector, along radii, towards the edge of the reflector. When applied to an 11 m diameter antenna, this method provided an rms repeatability of measurement of points on the reflector, averaged over the whole reflector surface, of 0.037 mm.

A disadvantage of these methods is that almost all of them must be carried out with the antenna pointing vertically upwards [Rusch, 1984]. However, an antenna is not usually operated in this position, and gravitational forces can cause the main reflector to deform when the antenna is pointing in another direction. One way of determining the deformation of a single point on the reflector is to measure the extension of a spring, one end of which is attached to the point, with the other end attached to the feed or subreflector via a taut length of wire [Anderson and Groth, 1963]. Another technique involves directing a laser beam, at a fixed angle from the antenna axis, towards the point on the reflector surface. Deformations of the surface can be deduced by utilizing two photocells to track the point at which the beam hits the surface [Slater, 1971]. These methods can also measure deformations caused by wind and temperature changes.

Some of the methods which have been developed for measuring the shapes of main reflectors are extremely tedious. For example, a full survey of 512 targets, employing pentaprisms, took 26 nights to complete [Slater, 1971]. However, a trolley designed to check surface curvature completed its measurements along 48 radii in only $3\frac{1}{2}$ hours [Payne *et al.*, 1976].

3.3.2 Near field scanning techniques

Near field scanning techniques have been developed over the last two decades. The basis of these techniques is to measure the amplitudes and phases of orthogonal com-

ponents of the field over a surface near to the antenna. The antenna's far field pattern can then be calculated from the measured near field distribution by utilizing an appropriate *near field to far field (NF-FF) transformation* [Rudge *et al.*, 1982, Sec. 8.5]. When the surface is a plane, cylinder or sphere, the mechanics involved with data gathering are convenient, and the NF-FF transformation is computationally practicable [Yaghjian, 1986]. The test facilities are usually indoors and typically accommodate antennas with maximum dimensions of less than about 15 m [Yaghjian, 1986, Fig. 1]. However, the methods can also be applied in principle at antenna sites and for larger antennas.

This section summarizes various near field scanning techniques. Note that measurement techniques which make use of approximations which are valid in the Fourier Fresnel region (which is part of the near field region) are discussed in Section 3.3.3.

The antenna whose radiation characteristics are to be measured is called the *test antenna* while a second, usually smaller antenna, employed to make the measurements, is called the *probe*. The probe is scanned over a *measurement surface* which is fixed in space relative to the test antenna. To make a measurement of a vector component of the field at a point on the surface, the probe is placed at the point, and the test antenna is excited at a particular frequency. A record is then made of the amplitude and phase of the signal appearing at the terminals of the probe. Since phase is a relative quantity, the phase of the signal fed to the test antenna is used as a reference. The particular vector component of the field that is measured is determined by the polarization characteristics and orientation of the probe. By reciprocity (Sec. 1.2.3), the same data can be alternatively obtained by transmitting radiation from the probe and measuring the output signal from the terminals of the test antenna.

To simplify the NF-FF transformation, the distance between the test antenna and the probe must be large enough for multiple reflections between the two antennas to have an insignificant effect on the measurements [Kummer and Gillespie, 1978]. Because the probe is in the near field region the angle subtended by the test antenna at the probe may be significant. Therefore, radiation from different parts of the test antenna may arrive at the probe from appreciably different angles, and be weighted according to the radiation pattern of the probe. Ideally, the probe should have an omnidirectional radiation pattern, so that radiation from all directions can have equal weight. In practice, the NF-FF transformation can be adjusted to correct for the probe's directionality provided the radiation characteristics of the probe are known [Paris *et al.*, 1978; Yaghjian, 1986].

An obvious measurement surface for a high gain reflector antenna is a plane, situated parallel and close to the aperture plane. Measurement of the field over this plane is called planar near field scanning. As explained in Section 2.1.3 the field over a plane can be Fourier transformed to obtain the far field. Furthermore, by following the method described in Section 3.2, the geometrical defects can be inferred from phase measurements of a single vector component of the aperture field. The latter strategy was employed by Repjar and Kremer [1982] for a 3.96 by 4.75 m reflector antenna.

In planar scanning, the antenna is typically stationary, while the probe is moved by a mechanism which scans an area larger than the aperture. To avoid multiple path reflections, the mechanism and supports are covered with microwave absorbers [Rudge *et al.*, 1982, p. 597]. The position of the probe must be known accurately. Movement of the probe in a direction perpendicular to the plane has the same effect on the measured field as a shape defect of the main reflector, because they both produce a path length change of the rays received by the probe. Therefore, the distance between

the probe and the plane should be maintained to at least the tolerance specified for the main reflector. Another requirement, which can be difficult to achieve for a large scanning area [Wood, 1987], is that the signal from the probe be transferred to the receiving equipment without introducing phase changes dependent upon the position of the probe.

Many large reflector antennas are able to rotate about two perpendicular axes, so that they can point in any direction towards the sky. This movement can be utilized in a spherical near field scanning system by fixing the probe's position (e.g. by attaching it to a nearby tower) and by rotating the test antenna. This arrangement avoids the need for extra mechanical equipment to move the test antenna relative to the probe. The distance between the probe and the test antenna is not important provided it is large enough to avoid multiple reflections between the antennas. Because the probe is always pointing towards the test antenna, the effects of unwanted radiation from other directions can be minimized by designing the probe antenna to exhibit nulls in these directions [Hansen and Larsen, 1984].

Although spherical scanning is mechanically simpler than planar scanning, the associated computations are more complicated, the NF-FF transformation being based on the spherical wave expansion method [Rudge *et al.*, 1982, Sec. 8.5.4]. Once the radiation pattern is calculated, it may be inverse Fourier transformed to obtain an estimate of the aperture field, allowing geometrical defects to be deduced. This procedure has been performed for a 1.5 m diameter reflector antenna by Rahmat-Samii and Lemanczyk [1988].

A compromise in both mechanical and computational complexity is cylindrical scanning. This can be achieved by moving the probe along a linear track and rotating the antenna about an axis which is parallel to the track. The NF-FF transformation uses cylindrical wave functions [Rudge *et al.*, 1982, Sec. 8.5.3]. It has been applied to a 15 m $\times$ 1.5 m antenna by Wood [1987].

The NF-FF transformations accord perfectly with Maxwell's equations (Sec. 1.1.2), provided measurements are made over an infinitely large plane or cylinder, or over a full sphere surrounding the antenna. In practice, however, only finite areas of these surfaces can be scanned. For spherical scanning, the portion of the sphere over which measurements are possible depends on how the antenna is mechanically supported (e.g. an earth station antenna cannot point vertically downwards). Therefore, the measurement surfaces are inevitably truncated. However, provided there are not too many missing data, these near field scanning techniques can lead to useful estimates of the far field pattern [Rudge *et al.*, 1982, p. 624]. Also, restricting the area of the surface over which measurements are made reduces the time required to make the measurements.

3.3.3 Measurements of the Fourier Fresnel and far fields

The most direct way of determining whether an antenna meets its far field pattern specifications is to measure its far field pattern. As intimated in Section 1.2.2 far field patterns describe the far field of an antenna. These patterns can be measured directly in the far field region, or alternatively, they can often be straightforwardly inferred from measurements made in the Fourier Fresnel region (Sec. 2.1.3.5). Measurements of only the amplitude of a vector component of the field and of both the amplitude and the phase of a vector component of the field are discussed in Sections 3.3.3.1 and 3.3.3.2 respectively.

3.3.3.1 Amplitude measurements

As discussed for the case of earth station antennas in Section 2.4.2.2, the specifications on the far field pattern of an antenna are typically expressed as an envelope under which the gain pattern must lie. The gain pattern can be straightforwardly computed from the far field amplitude patterns of three orthogonal components of the field (see Sec. 2.2.2). Since the radial component vanishes in the far field region (Sec. 2.1.2.2), the amplitude patterns of only two tangential orthogonal components of the far field are required to be measured. The measuring arrangement is similar to spherical near field scanning (Sec. 3.3.2), but with the separation between the test antenna and the probe exceeding the minimum far field distance (Sec. 1.2.1). For instance, for a 30 m diameter antenna, operating at 6 GHz, this distance is 36 km.

For arrangements in which the probe is closer than the minimum far field distance, the far field amplitude pattern can be inferred from measurement of the Fourier Fresnel amplitude pattern by employing what is called the defocusing technique, which is now described. The Fourier Fresnel pattern of an aperture antenna is related to the aperture field distribution by (2.39). The change in the phase of each vector component of the aperture field due to defocusing (if the main reflector is paraboloidal) is given approximately by (3.8). Combining these two equations gives the Fourier Fresnel pattern of a defocused paraboloidal antenna:

$$\begin{aligned}
\grave{E}_x(u,v) &= jw(u,v)\frac{e^{-jkR}}{R}\mathrm{FT}\left\{E_x(x,y)e^{-jk\frac{D^2\rho^2}{8R}+jk\Delta z\frac{2\rho^2}{1+(4f/D)^2}}\right\} \\
\grave{E}_y(u,v) &= jw(u,v)\frac{e^{-jkR}}{R}\mathrm{FT}\left\{E_y(x,y)e^{-jk\frac{D^2\rho^2}{8R}+jk\Delta z\frac{2\rho^2}{1+(4f/D)^2}}\right\} \qquad (3.9) \\
\grave{E}_z(u,v) &= \frac{-u}{w(u,v)}\grave{E}_x(u,v)+\frac{-v}{w(u,v)}\grave{E}_y(u,v)
\end{aligned}$$

where D is the aperture diameter, R is the distance between the aperture and the probe, and ρ is defined by (3.6). The quadratic phase terms on the right of (3.9) cancel when the axial displacement of the feed is [Chu, 1971]

$$\Delta z = \frac{1}{R}\left[f^2+\left(\frac{D}{4}\right)^2\right] \qquad (3.10)$$

When the above equation is satisfied, (3.9) reduces to (2.31), which gives the far field pattern of the antenna before it was defocused. Therefore, measurement of the far field pattern can be simulated in the Fourier Fresnel region by making measurements with the antenna defocused by the amount specified in (3.10). The antenna is then refocused by moving the feed back to its original position, and the measured pattern is assumed to describe the far field pattern. The quadratic aperture phase distribution due to defocusing is, however, not exact, and a physical optics analysis suggests that the optimum feed displacement is between 0.9 and 0.95 times the value obtained by (3.10) [Johnson *et al.*, 1973]. The defocusing technique becomes less accurate with decreasing R, which should therefore be greater than about $D^2/(4\lambda)$ [Rudge *et al.*, 1982, p. 637].

Typically, the measurement arrangement is such that the probe is the transmitting antenna, so that all the measured quantities (i.e. signal amplitude and direction of probe relative to test antenna boresight) can be recorded in the vicinity of the test antenna [Blake, 1984, p. 369]. Because of this, the probe is usually referred to as the *source*. Sources can be classified according to their locations, the most common of

which are terrestrial, airborne, cosmic, or aboard geostationary satellites. These four types of location are discussed in the following four paragraphs.

A terrestrial source is typically located on a tower in the far field (or Fourier Fresnel) region. The source is fixed so that it points towards the test antenna and scanning is achieved by rotating the test antenna. This arrangement has the advantage that the frequency and polarization of the source are able to be chosen to match the desired characteristics of the test antenna. It can be difficult to make accurate measurements, because of reflections from nearby buildings, hills and the ground [Blake, 1984, p. 370]. The effects of ground reflections may be reduced if a large valley lies between the test antenna and the source. They can also be reduced somewhat by positioning the source as high as possible, for example on a mountain.

The best way to reduce the effects of ground reflections is to ensure that the test antenna is always directed towards a high elevation. A method of achieving this is to employ an aircraft to fly the source above the test antenna, which is pointed towards a high elevation and kept stationary [Shanklin, 1955]. A tracking device is required to monitor the angular position of the source with respect to the test antenna. The strength of the field incident upon the test antenna from the source may not be constant because of the difficulty of constantly pointing the source at the test antenna and of maintaining a constant distance between the source and the test antenna. These variations can be measured by a second *reference antenna*, placed near to the test antenna, and always directed towards the source. The measured variations can then be used to normalize the signal from the test antenna, thus removing the effect of the variations [IEEE, 1979, Sec. 9]. This method is especially suited for test antennas which cannot be mechanically steered. For steerable antennas, it has the advantage that the antenna remains in one position, so that deflections of the main reflector due to gravity are constant throughout the measurement process.

Another method in which the test antenna points skywards, but which does not require tracking equipment, utilizes cosmic radio sources. A cosmic radio source, which moves through the sky in a predictable manner, is used in place of a transmitting probe. The test antenna is rotated in directions relative to the angular position of the source. An appropriate source is required to have an angular extent of less than 0.2 times the half power beamwidth of the test antenna so that it does not smear the structure of the far field pattern. The source is also required to be sufficiently strong to enable measurement of far field pattern levels of 60 dB or more below the main beam response. Unfortunately the smallest sources are not the strongest, so the two requirements cannot always be fulfilled simultaneously [Baars, 1973, Sec. VI]. However, a number of radio sources do have accurately known radiation characteristics and are suitable for determining the peak gain of an antenna [IEEE, 1979, Sec. 12.4; Baars, 1973].

Strong point-like sources can often be provided by geostationary satellites, which are therefore commonly employed to measure the far field pattern of earth station antennas [Miya, 1981, Sec. 5.7.2; CCIR, 1986b, Sec. 5]. The frequency transmitted by a satellite is usually within the operational frequency bandwidth of an earth station antenna, but is not always suited for other large antennas. A further advantage for earth station antennas is that its main beam and near in sidelobes are measured with the antenna at a typical operational elevation. Satellites drift about their nominal geostationary position in a predictable manner and this must be taken into account when scanning the test antenna. Green [1983] has described the use of a satellite to measure the amplitude pattern of a 64 m diameter radio telescope antenna, at a

frequency of 1.7 GHz. He conducted a detailed study of the expected sources of noise, and predicted that levels 58 dB below the main beam level could be measured with a signal to noise ratio of 10 dB. A dynamic range of at least 50 dB was obtained in practice. An important conclusion was that the measurements should be carried out at night to avoid the effects of thermal noise from the sun. In an interesting variant of the satellite technique, Levy *et al.* [1967] have measured the amplitude pattern of a 70 m antenna with a source transmitter located on the moon. Employing a 2.3 GHz signal, they obtained a dynamic range of 62 dB. The measurements by both Green and Levy *et al.* were performed without a separate reference antenna.

It is usually impossible to measure the amplitude pattern in all directions. Fortunately, it is often only necessary to observe a far field pattern over a limited portion of the measurement sphere. For example, to determine whether or not an earth station antenna pattern meets the CCIR specifications (Sec. 2.4.2.2), the pattern is only required along a finite curve, on the measurement sphere, corresponding to a portion of the geostationary orbit. Such a curve is called a *radiation pattern cut* [IEEE, 1984]. It is often considered adequate to make measurements of the amplitude pattern along two orthogonal cuts which intersect at the boresight direction [Blake, 1984, p. 365]. The angular extents of the cuts are determined by mechanical limitations on the rotation of the test antenna, or by interference from separate, terrestrially based, microwave systems.

The techniques described in the above paragraphs can often be straightforwardly extended to make measurements of the amplitude pattern over a finite area of the measurement sphere, thereby providing two-dimensional data. It is usually sufficient to make measurements at a set of sample points which are spread over the measurement sphere (Sec. 3.4.2.1). The antenna is typically scanned along several quasi-parallel cuts in a raster fashion, with measurements being made at sample points along each cut [e.g. Godwin *et al.*, 1986]. These measurements can show whether the amplitude pattern meets its specifications over a solid angle. By applying the algorithms described in Section 3.5 and Chapter 4 to a measured amplitude pattern, the distribution of a vector component of the aperture field can be estimated, thereby permitting the antenna's geometrical defects to be determined. These algorithms are called *phase retrieval* algorithms, because they can also be utilized to determine the far field phase pattern (remember that only the amplitude pattern is measured).

3.3.3.2 Complex holography

If equipment is available to measure the phase pattern as well as the amplitude pattern, the geometrical defects can be estimated in the following way: the measured amplitude and phase patterns of the copolar far (or Fourier Fresnel) field are combined to form a (complex) copolar radiation pattern. This pattern is then inverse Fourier transformed to provide an estimate of the copolar aperture field distribution (Sec. 2.1.3.4), which in turn is utilized to estimate the geometrical defects of the aperture in the manner described in Section 3.2. This method of estimating the geometrical defects is here called *complex holography* which is short for 'complex (amplitude and phase) microwave Fourier holographic metrology' [cf. Anderson, 1977]. Refer back to the introduction to this chapter for a discussion of the meaning of the word 'holography'.

The widespread use of complex holography is illustrated in Table 3.1, which gives details of complex holographic measurements made on several different antennas. To perform the measurements, the same kinds of sources as those used for amplitude

Antenna(s) and reference	Frequency GHz	Number of samples	Source	Approximate measurement time hours
13 m antennas in the 5 km telescope [Scott and Ryle, 1977]	15.4	17 × 17	cosmic	5
3.66 m paraboloidal reflector [Godwin *et al.*, 1978]	10	61 × 61	terrestrial	–
Chilbolton 25 m antenna [Anderson *et al.*, 1978]	10	101 × 101	terrestrial	3
Chilbolton 25 m antenna [Godwin *et al.*, 1981]	11.51	101 × 101	satellite	$3\frac{1}{2}$
25 m antennas in the Very Large Array [Napier *et al.*, 1983]	4.86	23 × 23	cosmic	–
Texas 4.9 m antenna [Mayer *et al.*, 1983]	86.16	83 × 83	terrestrial	3
NASA/JPL 64 m deep space network antenna [Rahmat-Samii, 1984; 1985]	2.28	11 × 11	cosmic	1
13.7 m Cassegrain antenna [Godwin *et al.*, 1985]	6.14	55 × 55	terrestrial	$1\frac{1}{2}$
Effelsberg 100 m telescope [Godwin *et al.*, 1986]	11.786	195 × 195	satellite	12
OTC Sydney-1 18 m satellite earth station antenna [Kalcina *et al.*, 1987]	4	64 × 64	satellite	4
NASA/JPL 64 m deep space network antenna [Rahmat-Samii, 1987]	11.45	189 × 189	satellite	–

Table 3.1 Details of several complex holographic measurements.

measurements (Sec. 3.3.3.1) can be utilized.

Because phase is not an absolute quantity, the phase of the signal from the test antenna must be compared to the phase of a reference signal. Unlike in near field scanning techniques (Sec. 3.3.2), it is impractical to compare the signal received by the test antenna with the signal fed to the source antenna, because of the large distance between these two antennas. However, a separate reference antenna can be utilized to provide a reference signal which is proportional to the field incident upon the test antenna. The reference antenna should therefore be situated close to the test antenna and be kept pointing towards the source antenna, while the test antenna moves relative to the source. If the source is stationary, the position of the reference antenna can be fixed. If the source moves then the reference antenna must be able to track the source, unless the source always remains well within the main beam of the reference antenna. When the only antenna occupying the site is the test antenna, a small, portable antenna is often utilized as the reference antenna [e.g. Godwin *et al.*, 1986]. Mayer *et al.* [1983] employ a horn mounted on the test antenna as a reference antenna. When the antennas in an array are to be measured, one of them can act as the reference antenna and measurements on all the others can be performed simultaneously [Scott and Ryle, 1977]. From (2.53), the power level of the reference signal depends on the relative polarization of the reference antenna and the radiation from the source. The reference signal power is greatest when the reference antenna and the source radiation are polarization matched.

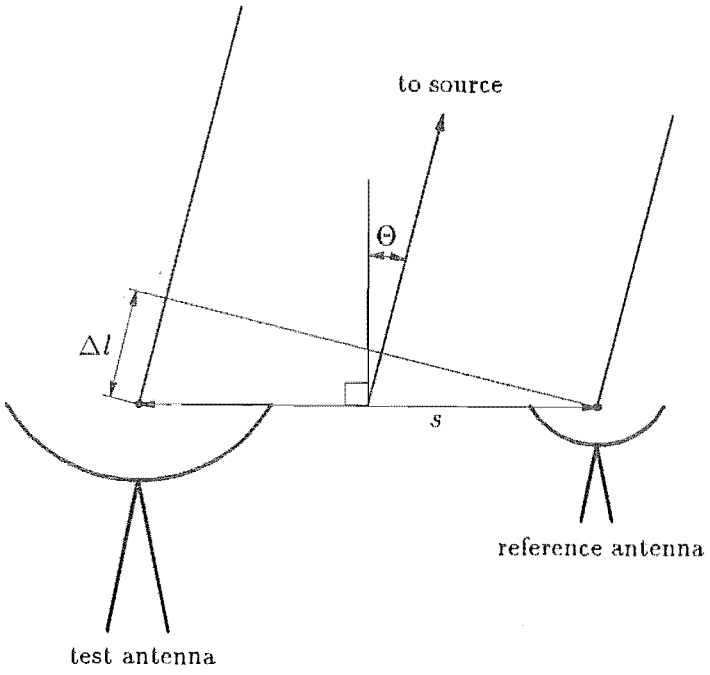
All of the sources discussed in Section 3.3.3.1, with the exception of terrestrial sources, move relative to the test antenna. Even geostationary satellite sources suffer minor perturbations about their nominal positions because of irregularities in the earth's gravitational field and gravitational forces from the sun and moon. [Mittra *et al.*, 1983, Sec. 1.4; Green, 1983]. Radial movements of the source, relative to the test antenna, do not affect the amplitude and phase pattern measurements because the field radiated by the source suffers the same change at both the test and reference antenna. However, any angular movement of the source does affect the phase measurements. This is because the phase of the copolar component of the incident field at the test antenna relative to the phase of the copolar component of the field at the reference antenna varies with the angular position of the source. From Figure 3.4(a), the phase variation $\Delta\Psi$ is given by

$$\Delta\Psi(\Theta) = -k\,\Delta l = -ks\sin\Theta \tag{3.11}$$

where Δl is the path length variation, s is the distance between the test and reference antennas, and Θ is the angle that the source makes with the plane comprising points equidistant from the test and reference antennas. Provided that the motion of the source is known, this phase variation can be subtracted from the measured phase. If the source's motion is unknown, but is almost periodic, which is the case for geostationary satellite and cosmic sources, the motion can be predicted by monitoring the phase difference between the signals from the test and reference antennas when they are both pointing towards the source. If the source is stationary, $\Delta\Psi$ is constant and can therefore be ignored for the reasons given in Section 3.2.2.

Antennas which can rotate usually do so about a point which is behind the main reflector. The development of the Fourier transform relationship (Sec. 2.1.3.2) between the aperture and far fields assumes that the measurements are made over a sphere centred on a point on the aperture plane. Rotating the antenna about a point behind the

(a)



(b)

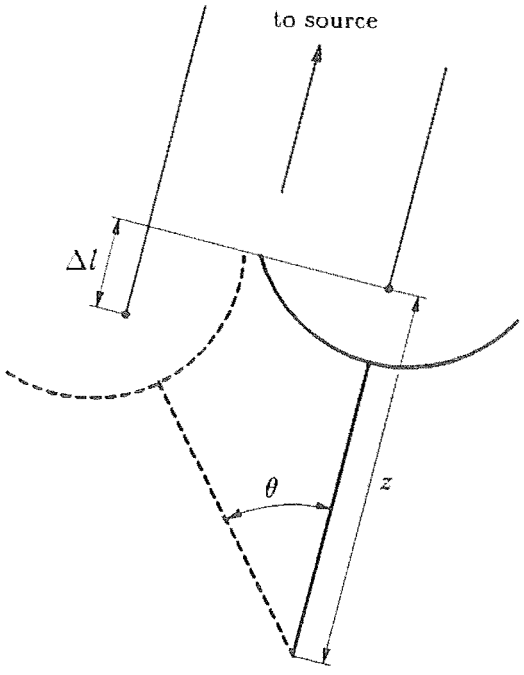


Figure 3.4 Corrections required for phase measurements in complex holography. Phase errors can be due to (a) motion of the source and (b) rotation of the test antenna about a point offset from the aperture plane.

aperture plane therefore introduces another phase variation, which from Figure 3.4(b) is [Scott and Ryle, 1977]

$$\Delta\Psi(\theta) = -k\,\Delta l = -kz(1-\cos\theta) \tag{3.12}$$

where the point of rotation is located a distance z behind the aperture plane and θ is the angle between the test antenna boresight and the direction of the source. By utilizing the relations in (2.28), and ignoring the constant phase term, (3.12) can be rewritten as

$$\Delta\Psi(u,v) = 2\pi z w(u,v) \tag{3.13}$$

The effect of the phase deviation can be removed by subtracting it from the measured phase pattern.

To calculate the aperture field using (2.34), it is assumed that the field over the whole radiation hemisphere is known. However, this field is fully determined by samples on grid in the u, v plane whose spacing satisfies the sampling theorem (Sec. 3.4.1.3). For test antennas which rotate in azimuth and elevation, Rahmat-Samii [1985, Appendix I] provides the relationship between the direction in which the antenna is pointing, specified as elevation and azimuth angles, and the corresponding point in the u, v plane. Under computer control, the antenna can be made to scan a regular grid of sample points in the u, v plane, allowing convenient use of the inverse FFT algorithm (Sec. 3.4.1.4) to compute the aperture field. Data which are measured at irregularly spaced sample points must be interpolated onto a regular grid before applying the inverse FFT algorithm [Rahmat-Samii and Cheung, 1987].

The inverse FFT algorithm computes an array of samples of the aperture field. As explained in Section 3.4.1.4, the area spanned by the samples in the u, v plane determines the spacing of the samples, and therefore the resolution, in the aperture plane. In particular, if the samples are distributed over the whole radiation hemisphere, the resolution in the aperture is $\lambda/2$. However, to achieve this, an extremely large number of samples are required: for example, a 30 m diameter antenna, operating at 6 GHz, requires over 1.5 million samples to fully determine the far field pattern over the whole radiation hemisphere. Fortunately, a coarser resolution in the aperture is usually adequate. Bennett and Godwin [1977] have performed computer simulations which indicate that to properly estimate displacements of the panels which comprise the main reflector, the resolution in the aperture need be no smaller than one quarter of the smallest panel dimension. Godwin *et al.* [1981] (see Table 3.1 for details) obtain this resolution from data extending to $\pm 2.4°$, from the direction of the source, in both azimuth and elevation. If one is only interested in feed displacement, 49 points on a square grid in the aperture plane are usually adequate, implying that measurements are only required over a correspondingly smaller angular range [Godwin *et al.*, 1978].

The method of determining geometrical defects, described in Section 3.2, makes use of the distribution of only one component of the aperture field. It is here assumed that the copolar component of the aperture field is utilized, although any other non-zero component can be used instead. Because the angular extents of the measurements are usually small, (2.57) can be utilized to describe the relationship between the aperture and far fields. The first equation of (2.57) reveals that only the copolar radiation pattern need be measured. This measurement can be performed by employing a source which is polarization matched to the copolar field of the test antenna.

If the measurements are made in the Fourier Fresnel region, the defocusing method discussed in Section 3.3.3.1 can be invoked to simulate the far field in the Fourier Fresnel

region. However, it is far simpler to perform a computational correction suggested by (2.58): a quadratic phase term is added to the inverse Fourier transform of the measured data to generate an estimate of the copolar aperture field distribution. If the copolar far field pattern is required, it can be computed from this copolar aperture field distribution. An informative discussion of this approach is provided by McGrane [1983].

A standard approach to determining the accuracy of the complex holographic method is to make two sets of measurements and then compute the rms difference between the geometrical defects inferred from each. Reported accuracies range from 0.1 mm [Godwin *et al.*, 1986] down to 0.004 mm [Mayer *et al.*, 1983].

Some of the references listed in Table 3.1 describe experiments which aim to assess the validity of the complex holography approach by employing it to estimate known geometrical defects. These known defects can be introduced by attaching conducting sheets to the main reflector [Mayer *et al.*, 1983; Kalcina *et al.*, 1987] or by purposely displacing one or more panels [Godwin *et al.*, 1986]. Other references report improvements in the test antenna performance by repositioning the feed [Godwin *et al.*, 1978], redesigning the subreflector [Godwin *et al.*, 1985] or realigning the main reflector panels [Godwin *et al.*, 1986] on the basis of information provided by complex holography.

When a main reflector exhibits shape defects which are directly behind a strut, the far field pattern tends to be unaffected in directions close to boresight, because radiation from these defects is largely blocked by the strut. Such defects do tend, however, to significantly affect the far field pattern at wide angles from boresight. Scattering from struts also affects the radiation at wide angles. Cook *et al.* [1985; 1987; 1989] have developed an extension of complex holography in which measurements of the complex vector far field pattern are made over a wide angle from the boresight and the measured data are processed to produce an estimate of the three-dimensional distribution of current throughout the antenna volume. Their approach is the same as that used by [Minard *et al.*, 1985, Sec. 4] who construct two-dimensional images of sonic scatterers by appropriate processing of data obtained from one-dimensional measurements of the far field of the scatterers. For radio antennas the theory can be conveniently summarized with the aid of the notation employed in Section 2.1.2. The far field of an antenna can be expressed as (cf. (2.23))[Silver, 1949, p. 88]

$$\mathbf{E}(\mathbf{r}) = \frac{-j\omega\mu}{4\pi r} e^{-jkr} \int_V [\mathbf{J} - (\mathbf{J}\cdot\hat{\mathbf{r}})\,\hat{\mathbf{r}}] e^{jk\mathbf{r}'\cdot\hat{\mathbf{r}}} dV \tag{3.14}$$

where $\mathbf{J}$ is the distribution of electric current throughout a volume V enclosing the antenna. Cook *et al.* [1989] conveniently ignore the second term in the square brackets. By employing the definitions in (2.28), the x component of the remaining term of $\mathbf{E}(\mathbf{r})$ can be expressed as

$$\begin{aligned} \acute{E}_x(u,v) &= \frac{-j\omega\mu}{4\pi R} e^{-jkR} \iiint J_x(x,y,z) e^{j2\pi(ux+vy+w(u,v)z)}\, dx\, dy\, dz \\ &= \frac{-j\omega\mu}{4\pi R} e^{-jkR} \int \mathrm{FT}\{J_x(x,y,z)\}\, e^{j2\pi w(u,v)z} dz \end{aligned} \tag{3.15}$$

where the Fourier transform operator is applied with respect to the variables x and y but not z. Consider the multiplication of $\acute{E}_x(u,v)$ by $e^{-j2\pi w(u,v)z_0}$ (cf. (3.13)). The inverse Fourier transform of the product is

$$\begin{aligned} &\mathrm{IFT}\left\{\acute{E}_x(u,v) e^{-j2\pi w(u,v)z_0}\right\} \\ &\quad = \frac{-j\omega\mu}{4\pi R} e^{-jkR} \int J_x(x,y,z) \odot \mathrm{IFT}\left\{e^{j2\pi w(u,v)(z-z_0)}\right\} dz \end{aligned} \tag{3.16}$$

where $\odot$ denotes convolution, which is defined in Table 3.3. When $z = z_0$ the integrand of (3.16) equals J_x over the plane $z = z_0$. When $z \neq z_0$ the integrand is a blurred form of J_x over the corresponding plane. Therefore, the operation defined by the left side of (3.16) produces a two-dimensional current distribution which is focused on the plane $z = z_0$. Similar results hold for the y component of the field. When (3.16) is repeatedly evaluated, with different z_0, an estimate of the three-dimensional current can be built up. By employing this technique, it is possible to infer the currents in regions of the main reflector which are behind the struts. Straightforward extension of the method described in Section 3.2 enables geometrical defects of these regions to be inferred.

3.3.4 Comparison of the measurement methods

In this section the different measurement techniques outlined in Sections 3.3.1 to 3.3.3 are compared with each other. They are first discussed in terms of their ability to provide an accurate estimate of the far field pattern for the purpose of determining whether of not an antenna meets its far field specifications. Then the various techniques are compared with respect to their overall effectiveness and convenience for estimating the locations and magnitudes of geometrical defects. The choice of which measurement technique is best for a given antenna depends upon what type of antenna it is, the required measurement accuracy and the availability of equipment and expertise. An important consideration is cost, which increases with the time needed to perform the measurements and with any extra equipment that must be employed. With computing power becoming continually cheaper, the expense of running the most demanding algorithms on a computer is here assumed to be a very small fraction of the total cost of performing the measurements.

An estimate of the radiation pattern can be determined in an indirect way when the geometry of the antenna is specified [IEEE, 1979, p. 62]. The geometry can sometimes be measured directly, or can be inferred from measurements of the field radiated by the antenna. This is particularly relevant when measurements of the radiation pattern are made at one frequency, with the radiation pattern being required to be estimated at another frequency [Godwin *et al.*, 1981; 1985]. The accuracy of the computed radiation pattern is limited by the accuracy to which the feed characteristics and geometry of the antenna are known and by the accuracy of the analysis technique that is employed.

Near field scanning techniques provide estimates of the far pattern through the computation of a NF-FF transformation. A disadvantage of the planar and cylindrical techniques is that they require equipment which moves the probe relative to the test antenna with great precision. However, an advantage of the planar technique is that, because the test antenna remains stationary, the geometrical deflections due to gravity remain fixed throughout the measurement [Mayer *et al.*, 1983]. Measurements of both the amplitude and phase of two orthogonal polarizations are required. The number of samples, and therefore the measurement time, is proportional to the square of the antenna's diameter in wavelengths [Yaghjian, 1986]. Therefore, for high gain antennas, the measurement time may be exceedingly large [Yaghjian, 1986].

Direct measurement of the Fourier Fresnel or far field tends to be the most rapid way of estimating the radiation pattern. It often requires less ancillary equipment than other methods. If possible, the test antenna is rotated about its own axes, thereby allowing the source to be stationary. Most reflector antennas have receivers which are able to measure the amplitude (or the power, which is proportional to the amplitude squared) of the received signal [Morris, 1985]. The measuring equipment needs a higher

dynamic range than for near field scanning, because in the far field region there tends to be a larger difference between the levels of the main lobe peak and the nulls than in the near field region [Jull, 1981, Sec. 4.3].

The most direct method of estimating the geometrical defects of an antenna is to measure its geometry. An important disadvantage of this method is, however, that most techniques for measuring the shape of the main reflector are more difficult, and take longer, to perform than, say, the measurements associated with complex holography [Scott and Ryle, 1977; Rahmat-Samii, 1984]. Another disadvantage is that many of the techniques require the antenna to be directed vertically upwards during the measurement, which may not be a typical operational position for the antenna. Ancillary equipment often has to be specially constructed to suit the particular antenna being measured.

The method, outlined in Section 3.2, of determining geometrical defects from the phase of the copolar aperture field distribution has an advantage over measuring the geometry directly, because it can also detect deviations in the field radiated by the feed. The planar near field scanning technique is the most direct method of estimating the phase distribution of the copolar aperture field. This technique has already been discussed in this section.

Complex holography requires equipment to measure both phase and amplitude, and a separate reference antenna. On many sites these are readily available. Depending on the chosen source, the measurements can be conducted with the antenna pointing at a typical operational angle, and therefore with typical gravitational deflections. Provided the measurements are made over a small solid angle, these deflections do not change significantly during the measurement process. A strong source is required to achieve the large dynamic range required to measure both the main beam and the nulls of the radiation pattern. Satellite sources are usually stronger than cosmic sources [Rahmat-Samii, 1985]. The signals from terrestrial sources can be still stronger, because they are closer to the test antenna.

The points made in the previous paragraph also apply to measurement of only the amplitude pattern, except that a reference antenna and phase measuring equipment are not then required. Estimating the geometrical defects using phase retrieval algorithms usually requires more computer processing than does complex holography. Furthermore, the algorithms described in Section 3.5 and Chapter 4 require measurements to be made at either two or four times as many sample points as required by complex holography (Sec. 3.4.2.1). Therefore, the use of phase retrieval algorithms is more appropriate than complex holography for antennas at sites which have no suitable reference antenna or no phase measuring equipment, and for which the cost of acquiring and employing this equipment outweighs the cost of recording the extra measurements. At frequencies over about 100 GHz, phase retrieval algorithms become even more important because the measurement of phase is difficult [McCormack and Anderson, 1988] due to the phase stability required in all components between the antennas and the measuring equipment. When details of the motion of the source are unknown, the phase retrieval approach, unlike complex holography, does not require the movement to be monitored. The significance of the phase retrieval approach can also be gauged from the many phase retrieval methods which have been proposed. They are discussed in Section 3.5.

3.4 PHASE RETRIEVAL FROM FOURIER TRANSFORM AMPLITUDE

In Section 3.3 it is explained how information about the geometrical defects of an antenna may be inferred from measurements of either the reflector profiles, the copolar aperture field phase distribution, or the copolar far field phase and amplitude patterns. However, it can be inconvenient, if not impossible, to measure phase directly, for reasons given in Section 3.3.4. On the other hand the measurement of an antenna's copolar amplitude pattern is one of the most straightforward radio engineering measurements that can be made (see Sec. 3.3.4). It would therefore be convenient to determine an antenna's geometrical defects from its copolar amplitude pattern. Since the geometrical defects can be inferred from the copolar aperture field phase distribution (Sec. 3.2), the problem reduces to retrieving this phase distribution from the measured copolar amplitude pattern. This is an example of the Fourier phase problem, which is discussed in this section.

The Fourier phase problem arises in diverse situations, examples of which are presented in Table 3.2. In this section it is discussed in non-specific terms, i.e. without particular reference to antenna engineering.

Consider a quantity, here called an *image*, represented by the complex scalar function $f(x, y)$. The real parameters x and y are conveniently thought of as Cartesian coordinates in the *image plane*. The *Fourier transform* of the image is denoted by $F(u, v) = \mathrm{FT}\{f(x, y)\}$ (defined in Table 3.3), which is also a complex scalar function. The real parameters u and v are considered to be Cartesian coordinates in the *Fourier plane*. The *Fourier phase problem*, which Bates and McDonnell [1989, Sec. 20] specify

Application	Image	Fourier transform	Causes
electron microscopy	transmissivity or reflectivity of the specimen (complex)	back focal plane field	frequencies of the field are too high to measure [Misell, 1978]
acoustic microscopy	reflectivity of the specimen (complex)	response of microscope	when diode detecting the microscope's response, its phase is lost [Fright *et al.*, 1989]
astronomy	spatially incoherent radiating source distribution (real and positive)	visibility	the phase of the field is distorted by the atmosphere [Bates, 1982] and instrumental effects [Pearson and Readhead, 1984]
speech transmission	speech signal (real)	temporal spectrum	discard phase to increase transmission efficiency as in LPC coding of signals [Makhoul, 1975]

Table 3.2 Examples of applications in which the Fourier phase problem occurs [Lane, 1988].

in a more general context, is here posed as

$$\begin{gathered}\text{given } |F(u,v)|, \\ \text{retrieve the image } f(x,y)\end{gathered} \tag{3.17}$$

It is called a 'phase' problem, because once phase$\{F(u,v)\}$ is known, $f(x,y)$ can be immediately calculated from IFT$\{F(u,v)\}$.

In most situations, more than just $|F(u,v)|$ is known. This knowledge may relate to properties of the image, or can be partial quantitative information concerning either the Fourier transform phase or the image. This additional information can be incorporated into the procedure invoked to solve the Fourier phase problem, or it can be used to verify the solution.

Practical computational aspects of solving the Fourier phase problem are discussed in Section 3.4.1. In Section 3.4.2, the uniqueness of the Fourier phase problem, when no additional information is available, is discussed. Section 3.4.3 presents a practical algorithm for solving the Fourier phase problem, when additional information of the kind intimated in the previous paragraph is available. Throughout this thesis, an image is denoted by a lower case roman letter (e.g. $f(x,y)$), while its Fourier transform is denoted by the same letter in upper case roman (e.g. $F(u,v)$).

3.4.1 Computer processing details

Before describing the details (in Secs. 3.4.3 and 3.5) of practical algorithms for solving the Fourier phase problem it is convenient in this section to discuss several computational considerations which are critical for these algorithms. Estimates of the image and its Fourier transform are stored in arrays, rather than as functions. This means that the computer must operate on samples of the image and its Fourier transform, and, where necessary, employ a discrete form of the Fourier transform operator. The way in which these discrete images and operations relate to their continuous analogues is analysed in the following sections. Although the discussion in the next three sections is in terms of the image $f(x,y)$, the concepts (e.g. compactness and sampling) also apply to the Fourier transform $F(u,v)$. Functions which are not defined in the text are listed in Table 3.3, as are relevant properties of Fourier transformation.

3.4.1.1 Compact images

An important property of an image is its *energy* which is defined to be

$$\text{Energy of } f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |f(x,y)|^2\,dx\,dy \tag{3.18}$$

Most images of interest contain finite energy. For an image $|f(x,y)|$ to have finite energy it must tend to zero as either $|x|$ and/or $|y|$ tend to infinity. Therefore, most of its energy is concentrated over a finite region of space. In practice, no measurement technique can detect parts of the image with an amplitude of less than some positive real number, say ξ, which is determined partly by the resolution of the detection method and partly by the noise which invariably contaminates the image. The *support* S^f of $f(x,y)$ is defined to be the region of the x,y plane for which $|f(x,y)| > \xi$. This means that the image is negligible outside its support.

The physical size of a support is defined by its extents, which are themselves defined in Figure 3.5. The *extents* L_x^f and L_y^f of $f(x,y)$, in the x and y directions respectively,

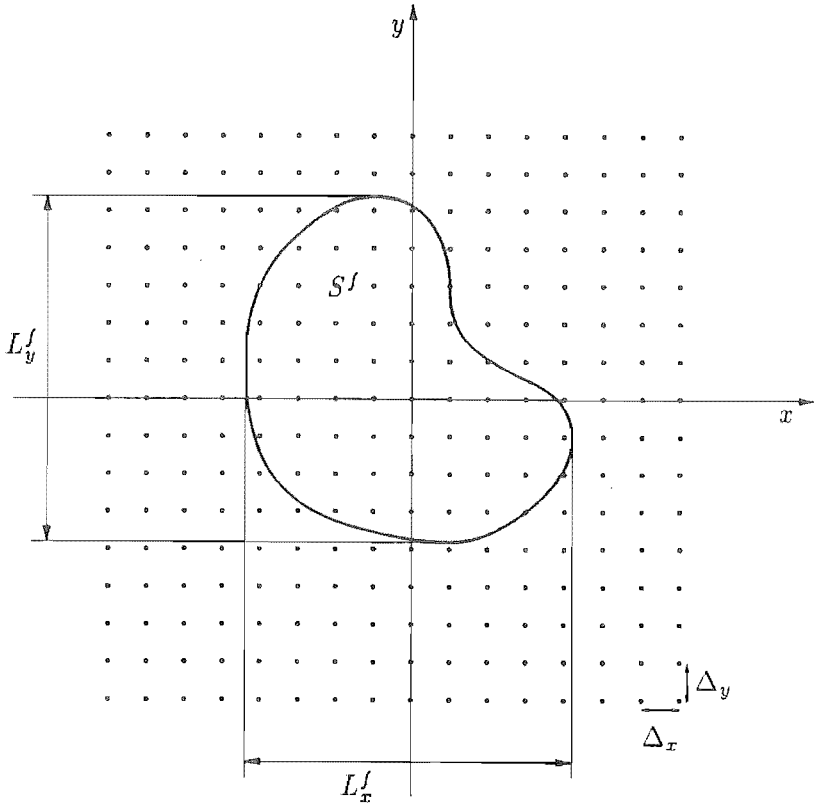


Figure 3.5 Support and extents of a compact image $f(x, y)$. Also shown is a grid of sampling points.

are the lengths of the rectangle (with sides parallel to the x and y axes) which just encloses S^f.

For a given value of ξ, an image is defined to be *compact* if its extents are both finite and if its amplitude is everywhere finite. A compact image is said to be *exactly compact* if $\xi = 0$ and *approximately compact* otherwise.

The *energy conservation theorem* for Fourier transforms (also called Rayleigh's theorem [Bracewell, 1978, p. 112]) states that the energy of $f(x, y)$ is equal to the energy of its Fourier transform $F(u, v)$ [Bates and McDonnell, 1989, p. 24]. Therefore, if $f(x, y)$ contains finite energy, $F(u, v)$ also contains finite energy. A consequence of this is that both $F(u, v)$ and $f(x, y)$ are compact. However, it is not possible for both of them to be exactly compact [Slepian, 1983]. It often simplifies the mathematics (e.g. in Sec. 3.4.1.4) to assume that parts of an image having amplitudes less than ξ are in fact identically equal to zero. This is equivalent to assuming that an approximately compact image is exactly compact.

3.4.1.2 Sampling

A digital computer can store and manipulate only a finite amount of discrete data. The data representing an image are usually in the form of samples on a rectangular grid. Figure 3.5 shows a grid of M by N sample points, with a sample spacing of Δ_x and Δ_y,

in the x and y directions respectively. For compatibility with the fast Fourier transform algorithm (Sec. 3.4.1.4), both N and M are taken to be integer powers of two. It is assumed that each sample is the value of the image averaged over an infinitely small area [Bates and McDonnell, 1989, Sec. 11].

A finite set of samples representing an image $f(x,y)$ is referred to as a *sampled image* and is denoted by $f[m,n]$, where the integers m and n are indices defining the position of the sample points. Expressed mathematically, the sampling process is described by the Samp $\{\cdot\}$ operator, which is defined by

$$\begin{aligned} f[m,n] &= \text{Samp}\,\{f(x,y),\Delta_x,\Delta_y,M,N\} \\ &= f(m\Delta_x,n\Delta_y) \\ &\qquad \text{for } \tfrac{-M}{2} \le m \le \tfrac{M}{2}-1,\ \tfrac{-N}{2} \le n \le \tfrac{N}{2}-1 \end{aligned} \tag{3.19}$$

Whenever a sampled image is introduced in this thesis, it is implicitly assumed that the sample spacing and number of samples in each direction have already been specified.

The sample points span a finite area of the x,y plane. The *resolution cell* represented by a single sample point is a rectangle of area Δ_x by Δ_y. Therefore, the area spanned by a grid of M by N samples is taken to be a rectangle whose sides are of length $M\Delta_x$ and $N\Delta_y$ in the x and y directions respectively. To ensure that the sample points span at least the whole of the support of the image, it is required that

$$M\Delta_x \ge L_x^f \quad \text{and} \quad N\Delta_y \ge L_y^f \tag{3.20}$$

Failing to meet these requirements is equivalent to truncating the image.

Although it cannot be stored on a computer, it is convenient to consider a sampled image for which M and N are infinitely large. An alternative way of expressing the connection between $f(x,y)$ and its samples $f[m,n]$ is then [Bracewell, 1978, Chap. 10]

$$f(x,y)\,\text{III}(\frac{x}{\Delta_x},\frac{y}{\Delta_y}) = \Delta_x\Delta_y \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} f[m,n]\,\delta(x-m\Delta_x,y-n\Delta_y) \tag{3.21}$$

where both the delta function $\delta(x,y)$ and the grid of deltas $\text{III}(x,y)$ are defined in Table 3.3.

3.4.1.3 Interpolation and aliasing

It is sometimes possible to reconstruct an image exactly from its samples (assuming that the samples are free of noise and any other source of uncertainty). Consider an image $f(x,y)$ whose Fourier transform $F(u,v)$ is exactly compact, with extents L_u^F and L_v^F. For sample spacings of Δ_x and Δ_y in the x and y directions respectively, it is convenient to introduce *sampling factors* α_x and α_y defined by

$$\alpha_x\Delta_x = \frac{1}{L_u^F} \quad \text{and} \quad \alpha_y\Delta_y = \frac{1}{L_v^F} \tag{3.22}$$

The quantities $1/L_u^F$ and $1/L_v^F$ are the *Nyquist sample spacings* in the x and y directions respectively [Brigham, 1974, Sec. 5-4]. The *sampling theorem* [Bracewell, 1978, Chap. 10; Bates and McDonnell, 1989, Sec. 10] states that $f(x,y)$ can be exactly recovered from its samples, provided that both M and N are infinitely large, and that neither sampling factor is less than unity. For compact images, M and N need only

Operation or function	Definition	Pages
Fourier transformation	$F(u,v) = \mathrm{FT}\{f(x,y)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{j2\pi(ux+yv)}\,du\,dv$	[241,22]
inverse Fourier transformation	$f(x,y) = \mathrm{IFT}\{F(u,v)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(u,v)e^{-j2\pi(ux+yv)}\,dx\,dy$	[241,22]
convolution	$f(x,y) \odot g(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\xi,\eta)g(x-\xi,y-\eta)\,d\xi\,d\eta$	[243,24]
conjugate reflection	$\tilde{f}(x,y) = f^*(-x,-y)$	—
autocorrelation	$ff(x,y) = f(x,y) \odot \tilde{f}(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\xi,\eta)f^*(\xi-x,\eta-y)\,d\xi\,d\eta$	[115,25]
rectangle function	$\mathrm{rect}(x,y) = \begin{cases} 1 & \frac{-1}{2} \le x < \frac{1}{2},\ \frac{-1}{2} \le y < \frac{1}{2} \\ 0 & \text{elsewhere} \end{cases}$	[52,36]
sinc function	$\mathrm{sinc}(x,y) = \frac{\sin(\pi x)}{\pi x}\frac{\sin(\pi y)}{\pi y}$	[62,20]
delta function	$\delta(x,y) = 0$ for $(x,y) \neq (0,0)$, and $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\delta(x,y)\,dx\,dy = 1$	[69,23]
grid of deltas	$\mathrm{III}\left(\frac{x}{\xi},\frac{y}{\eta}\right) = \lvert\xi\eta\rvert \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\delta(x-m\xi,y-n\eta)$	[77,–]
sifting property	$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\delta(x-\xi,y-\eta)f(x,y)\,dx\,dy = f(\xi,\eta)$	[74,36]

Property or operation	Image	Fourier transform	Pages
convolution	$f(x,y) \odot g(x,y)$	$F(u,v)G(u,v)$	[243,24]
multiplication	$f(x,y)g(x,y)$	$F(u,v) \odot G(u,v)$	—
conjugate reflection	$\tilde{f}(x,y)$	$F^*(u,v)$	—
autocorrelation	$ff(x,y)$	$F(u,v)F^*(u,v) = \lvert F(u,v)\rvert^2$	[115,25]
translation	$f(x-\xi,y-\eta)$	$F(u,v)e^{j2\pi(u\xi+v\eta)}$	[104,–]
scalar multiplication	$cf(x,y)$	$cF(u,v)$	—
similarity	$f(\xi x,\eta y)$	$\frac{1}{\lvert\xi\eta\rvert}F\left(\frac{u}{\xi},\frac{v}{\eta}\right)$	[102,–]
rectangle function	$\mathrm{rect}(x,y)$	$\mathrm{sinc}(u,v)$	[389,–]
sinc function	$\mathrm{sinc}(x,y)$	$\mathrm{rect}(u,v)$	[389,–]
grid of deltas	$\mathrm{III}(x,y)$	$\mathrm{III}(u,v)$	[388,–]

Table 3.3 Definitions and properties related to Fourier transformation (the first and second numbers in the last column refer to the relevant pages in Bracewell [1978] and Bates and McDonnell [1989] respectively).

satisfy (3.20). The image can then be reconstructed by sinc interpolating its samples [Bates and McDonnell, 1989, Sec. 11]:

$$f(x,y) = \sum_{m=-M/2}^{M/2-1} \sum_{m=-N/2}^{N/2-1} f[m,n]\,\mathrm{sinc}(\frac{x}{\Delta_x} - m, \frac{y}{\Delta_y} - n) \tag{3.23}$$

where the sinc(·) function is defined in Table 3.3.

An image is said to be *oversampled* when both of the sampling factors α_x and α_y are greater than unity. Conversely, an image is *undersampled* when either of the sampling factors is less than unity. Undersampling implies that the samples are spaced too far apart and therefore do not adequately represent the continuous image. This effect is known as *aliasing* [Bracewell, 1978, Chap. 10]. The remainder of this section explains aliasing, with the aid of the one-dimensional example illustrated by Figure 3.6.

Consider an image $f(x,y)$ and its Fourier transform $F(u,v)$, as illustrated in Figure 3.6(a). Figure 3.6(b) depicts a grid of deltas spaced by Δ_x and Δ_y in the x and y directions respectively. The product of the image and the grid of deltas is shown in Figure 3.6(c) and is given by (3.21). The Fourier transform of (3.21) is

$$\begin{aligned} \mathrm{FT}\left\{ f(x,y)\,\mathrm{III}(\frac{x}{\Delta_x}, \frac{y}{\Delta_y}) \right\} &= \Delta_x \Delta_y F(u,v) \odot \mathrm{III}(u\Delta_x, v\Delta_y) \\ &= \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} F(u - \frac{m}{\Delta_x}, v - \frac{n}{\Delta_y}) \end{aligned} \tag{3.24}$$

where both the Fourier transform $\mathrm{FT}\{\cdot\}$ and the convolution $\odot$ operators are defined in Table 3.3. The Fourier transform in (3.24) is a periodic function created by the superposition of an infinite number of translated versions of $F(u,v)$ [Goodman, 1968, Sec. 2-3]. If $f(x)$ is undersampled (i.e. $\alpha_x, \alpha_y < 1$ in (3.22)) portions of one shifted version of $F(u,v)$ overlap onto other versions of it, as is depicted in Figure 3.6(c).

The two-dimensional period of the Fourier transform in (3.24) is $1/\Delta_x$ and $1/\Delta_y$ in the u and v directions respectively. The right hand side of (3.24) within a single period is here called the *alias* $A^F(u,v)$ of $F(u,v)$. The Alias $\{\cdot\}$ operator is defined by

$$\begin{aligned} A^F(u,v) &= \mathrm{Alias}\left\{ F(u,v), \frac{1}{\Delta_x}, \frac{1}{\Delta_y} \right\} \\ &= \Delta_x \Delta_y\, \mathrm{rect}(\Delta_x u, \Delta_y v)[F(u,v) \odot \mathrm{III}(\Delta_x u, \Delta_y v)] \\ &= \begin{cases} \displaystyle\sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} F\left(u - \frac{m}{\Delta_y}, v - \frac{n}{\Delta_y}\right) \\ \qquad \text{where } \frac{-1}{2\Delta_x} \le u < \frac{1}{2\Delta_x},\ \frac{-1}{2\Delta_y} \le v < \frac{1}{2\Delta_y} \\ 0 \qquad \text{elsewhere} \end{cases} \end{aligned} \tag{3.25}$$

where the rect(·) function is defined in Table 3.3 The quantities $1/\Delta_x$ and $1/\Delta_y$ are the *widths* of the alias. In the u direction, any non-zero portions of $F(u,v)$ which lie outside of the range $-1/(2\Delta_x) \le u < 1/(2\Delta_x)$ are translated by a multiple of $1/\Delta_x$ until they do lie within that range, as illustrated by Figure 3.6(d). Similar translations can occur in the v direction. The resulting translated portions of $F(u,v)$ are summed to form $A^F(u,v)$. This process is called *aliasing* [Bracewell, 1978, Chap. 10].

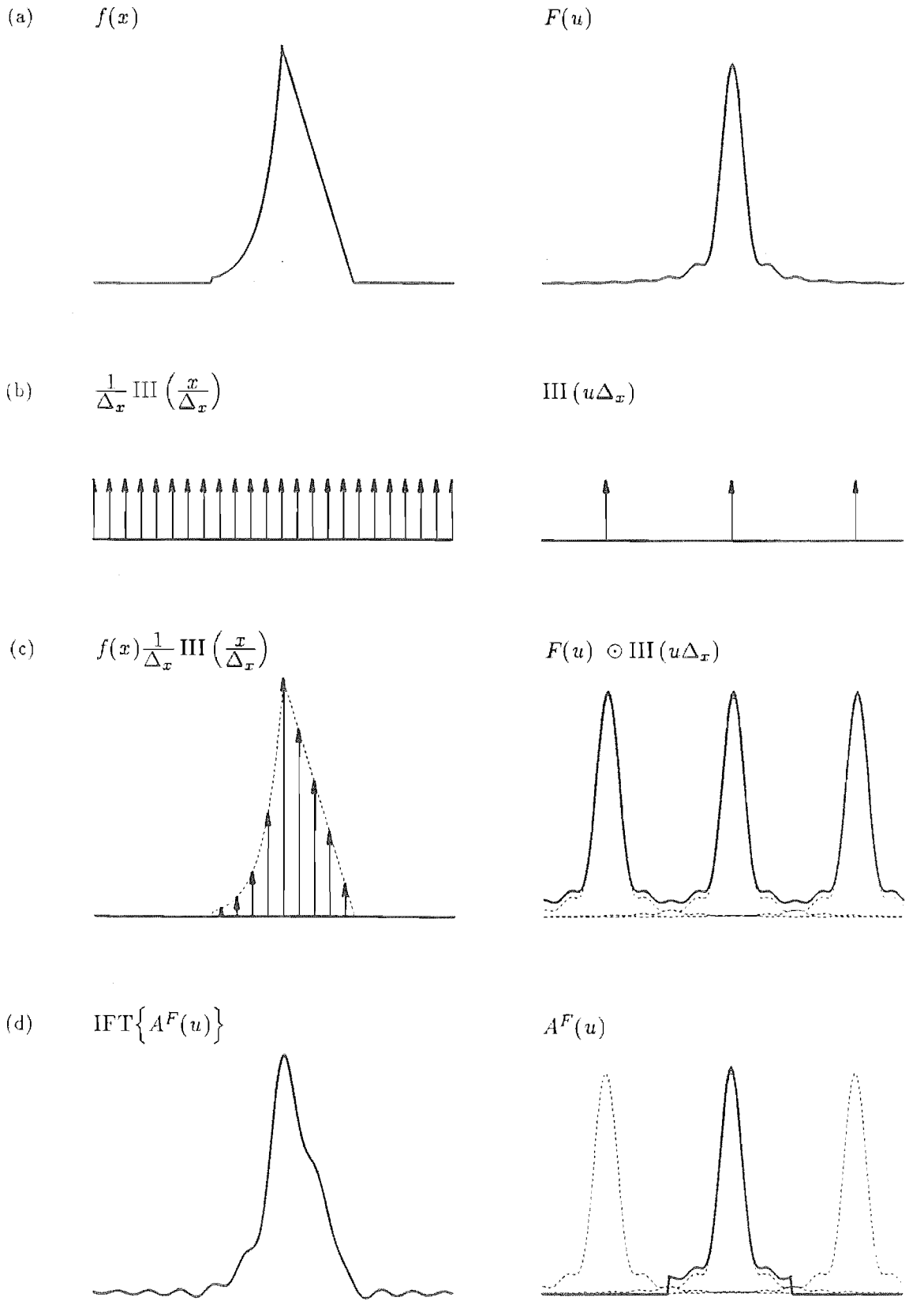


Figure 3.6 One-dimensional example of aliasing. Images are on the left and their corresponding Fourier transforms are on the right. The text explains the progression from (a) the image and its Fourier transform, through to (d) the image reconstructed from its samples and the alias of the Fourier transform. Vertical arrows represent delta functions.

The inverse Fourier transform of $A^F(u,v)$ is (see also Fig. 3.6(d))

$$\begin{aligned}\mathrm{IFT}\Big\{A^F(u,v)\Big\} &= \mathrm{IFT}\{\Delta_x\Delta_y\,\mathrm{rect}(\Delta_x u,\Delta_y v)[F(u,v)\odot \mathrm{III}(\Delta_x u,\Delta_y v)]\}\\ &= \frac{1}{\Delta_x\Delta_y}\,\mathrm{sinc}(\frac{x}{\Delta_x},\frac{y}{\Delta_y})\odot[f(x,y)\,\mathrm{III}(\frac{x}{\Delta_x},\frac{y}{\Delta_y})] \qquad (3.26)\\ &= \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} f[m,n]\,\mathrm{sinc}(\frac{x}{\Delta_x}-m,\frac{y}{\Delta_y}-n)\end{aligned}$$

which is equivalent to (3.23) when the image $f(x,y)$ is compact. If $F(u,v)$ is exactly compact, with each extent less than the corresponding alias width, then $A^F(u,v) = F(u,v)$, and therefore $\mathrm{IFT}\Big\{A^F(u,v)\Big\} = f(x,y)$. However, from (3.22), this can only occur if the image is oversampled by a factor of at least one. When the image is undersampled, aliasing occurs in the Fourier plane and $A^F(u,v) \neq F(u,v)$, implying that sinc interpolation of the image samples does not equal $f(x,y)$. This of course accords with the sampling theorem. Note that although $F(u,v)$ may not, in general, be exactly compact, its alias is always exactly compact.

3.4.1.4 The discrete Fourier transform (DFT)

The previous two sections show how to obtain a sampled representation of an image, suitable for processing by a digital computer. In the algorithms introduced in Section 3.4.3, one of the main operations performed on images is Fourier transformation. However, a sampled image cannot be Fourier transformed by a conventional Fourier transform operator: the discrete Fourier transform (DFT) operator must be utilized instead. This section derives the DFT operator from the Fourier transform operator, thereby showing their interrelationship. Figures 3.6 and 3.7 depict a one-dimensional example of the steps involved.

Figure 3.6(c) depicts an image multiplied by a grid of delta functions. From (3.24), the Fourier transform of this product is a periodic function. Figure 3.7(a) shows a grid of delta functions spaced by Δ_u and Δ_v in the u and v directions respectively. The product of the periodic Fourier transform and the grid of delta functions is illustrated in Figure 3.7(b) and its inverse Fourier transform is

$$\begin{aligned}&\mathrm{IFT}\Big\{\Delta_x\Delta_y[F(u,v)\odot \mathrm{III}(u\Delta_x,v\Delta_y)]\,\mathrm{III}(\frac{u}{\Delta_u},\frac{v}{\Delta_v})\Big\}\\ &\quad= \Delta_u\Delta_v[f(x,y)\,\mathrm{III}(\frac{x}{\Delta_x},\frac{y}{\Delta_y})]\odot \mathrm{III}(x\Delta_u,y\Delta_v)\end{aligned} \qquad (3.27)$$

which is itself periodic. Figure 3.7(b) should be compared with Figure 3.7(c), which depicts $A^J(x,y) = \mathrm{Alias}\{f(x,y),1/\Delta_u,1/\Delta_v\}$. One period of the function in (3.27) is given by $A^J(x,y)\,\mathrm{III}(x/\Delta_x,y/\Delta_y)$ provided that

$$M\Delta_x = \frac{1}{\Delta_u} \quad \text{and} \quad N\Delta_y = \frac{1}{\Delta_v} \qquad (3.28)$$

where M and N are integers. The Fourier transform of (3.27) can now be rewritten as

$$\begin{aligned}&\Delta_x\Delta_y[F(u,v)\odot \mathrm{III}(u\Delta_x,v\Delta_y)]\,\mathrm{III}(\frac{u}{\Delta_u},\frac{v}{\Delta_v})\\ &\quad= \mathrm{FT}\Big\{\Delta_u\Delta_v[A^J(x,y)\,\mathrm{III}(\frac{x}{\Delta_x},\frac{y}{\Delta_y})]\odot \mathrm{III}(x\Delta_u,y\Delta_v)]\Big\} \qquad (3.29)\\ &\quad= \mathrm{FT}\Big\{A^J(x,y)\,\mathrm{III}(\frac{x}{\Delta_x},\frac{y}{\Delta_y})\Big\}\,\mathrm{III}(\frac{u}{\Delta_u},\frac{v}{\Delta_v})\end{aligned}$$

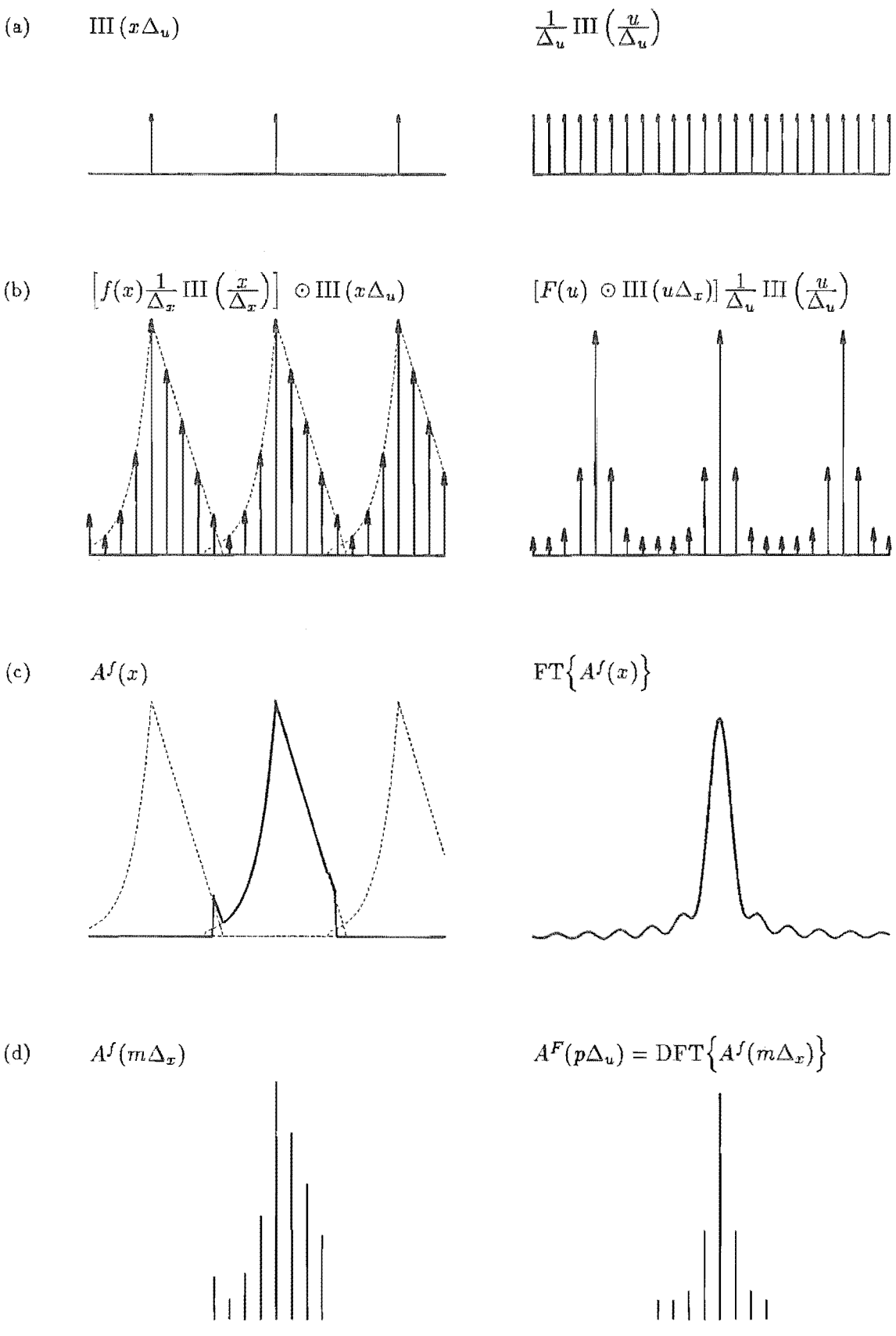


Figure 3.7 One-dimensional example of the relationship between the Fourier transform and DFT operators. This figure continues on from Figure 3.6(a) to (c). Images are on the left and their corresponding Fourier transforms are on the right. See text for explanation.

The coefficients of the centre M by N delta functions, corresponding to $\mathrm{III}(u/\Delta_u, v/\Delta_v)$ on the left of (3.29), are samples of $A^F(u,v) = \mathrm{Alias}\{F(u,v), 1/\Delta_x, 1/\Delta_y\}$. The samples are only defined at points $(u,v) = (p\Delta_u, q\Delta_v)$, where p and q are integers. Equating these coefficients with the coefficients of the delta functions on the right of (3.29), and employing the definition of the $\mathrm{FT}\{\cdot\}$ operator presented in Table 3.3, yields

$$\begin{aligned} &A^F(p\Delta_u, q\Delta_v) \\ &\quad = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A^f(x,y)\,\mathrm{III}(\frac{x}{\Delta_x}, \frac{y}{\Delta_y}) e^{j2\pi(xp\Delta_u + yq\Delta_v)}\,dx\,dy \end{aligned} \tag{3.30}$$

Knowing that $A^f(x,y)$ is exactly compact (Sec. 3.4.1.3), and invoking the definition of $\mathrm{III}(\cdot)$ and the properties of the delta function given in Table 3.3, the integral in (3.30) evaluates to

$$A^F(p\Delta_u, q\Delta_v) = \Delta_x\Delta_y \sum_{m=-M/2}^{M/2-1} \sum_{m=-N/2}^{N/2-1} A^f(m\Delta_x, n\Delta_y) e^{j2\pi(mp/M + nq/N)} \tag{3.31}$$

which relates samples defined in the image plane to samples defined in the Fourier plane and is depicted in Figure 3.7(d).

Let $g(x,y)$ be any image and $G(u,v)$ its Fourier transform. Define $g[m,n]$ and $G[p,q]$ by

$$\begin{aligned} g[m,n] &= \mathrm{Samp}\{A^g(x,y), \Delta_x, \Delta_y, M, N\} \\ G[p,q] &= \mathrm{Samp}\left\{A^G(u,v), \Delta_u, \Delta_v, M, N\right\} \\ \text{where}\quad A^g(x,y) &= \mathrm{Alias}\{g(x,y), M\Delta_x, N\Delta_y\} \\ A^G(u,v) &= \mathrm{Alias}\{G(u,v), M\Delta_u, N\Delta_v\} \end{aligned} \tag{3.32}$$

remembering the constraint (3.28). Then the *discrete Fourier transform (DFT)* $G[p,q]$ of $g[m,n]$, is defined to be [cf. Bracewell, 1978, Chap. 18; Bates and McDonnell, 1989, Sec. 12]

$$G[p,q] = \mathrm{DFT}\{g[m,n]\} = \Delta_x\Delta_y \sum_{m=-M/2}^{M/2-1} \sum_{m=-N/2}^{N/2-1} g[m,n] e^{j2\pi(mp/M + nq/N)} \tag{3.33}$$

which, it should be noted, accords with (3.31). It is similarly possible to derive from (3.27) the *inverse discrete Fourier transform* $g[m,n]$ of $G[p,q]$, which is defined to be [cf. Bracewell, 1978, Chap. 18; Bates and McDonnell, 1989, Sec. 12]

$$g[m,n] = \mathrm{IDFT}\{G[p,q]\} = \Delta_u\Delta_v \sum_{p=-M/2}^{M/2-1} \sum_{q=-N/2}^{N/2-1} G[p,q] e^{-j2\pi(mp/M + nq/N)} \tag{3.34}$$

Often, however, the image $f(x,y)$, rather than its alias, is sampled. To show the relationship between the DFT of this sampled image and the Fourier transform $F(u,v)$, let $g(x,y) = \mathrm{rect}(x\Delta_u, y\Delta_v) f(x,y)$ [Brigham, 1974, Sec. 6-4]. Then $A^g(x,y) = g(x,y)$ so that (3.32) becomes

$$\begin{aligned} g[m,n] &= \mathrm{Samp}\{f(x,y), \Delta_x, \Delta_y, M, N\} \\ G[p,q] &= \mathrm{Samp}\left\{A^G(u,v), \Delta_u, \Delta_v, M, N\right\} \\ \text{where}\quad A^G(u,v) &= \mathrm{Alias}\{G(u,v), M\Delta_u, N\Delta_v\} \\ G(u,v) &= \frac{1}{\Delta_u\Delta_v}\,\mathrm{sinc}\left(\frac{u}{\Delta_u}, \frac{v}{\Delta_v}\right) \odot F(u,v) \end{aligned} \tag{3.35}$$

The extents of the convolution of two quantities can usually (i.e. except in special cases) be expected to be larger than the extents of either of them. The convolution in the fourth equation of (3.35) implies that the extents of $G(u,v)$ are larger than those of $F(u,v)$. The effect of this is called *leakage* [Bergland, 1969].

For an approximately compact image $f(x,y)$ (see Sec. 3.4.1.1), whose Fourier transform $F(u,v)$ is also approximately compact, both leakage and aliasing can be avoided if the sample spacings, and the number of samples, are chosen with care. It is assumed that both $f(x,y)$ and $F(u,v)$ are exactly compact. Leakage is avoided if (3.20) holds, implying that $\text{rect}(x\Delta_u, y\Delta_v)f(x,y) = f(x,y)$ and, therefore, that $G(u,v) = F(u,v)$ in (3.35). This also avoids aliasing in the image plane. Aliasing in the Fourier plane is avoided by oversampling the image (as described in Sec. 3.4.1.3). Equation (3.32) can then be rewritten as

$$\begin{aligned} g[m,n] &= \text{Samp}\{f(x,y), \Delta_x, \Delta_y, M, N\} \\ G[p,q] &= \text{Samp}\{F(u,v), \Delta_u, \Delta_v, M, N\} \end{aligned} \tag{3.36}$$

with the requirements (3.28), (3.20) and (3.22), which are now repeated for emphasis:

$$\begin{aligned} M\Delta_x = \frac{1}{\Delta_u} \quad &\text{and} \quad N\Delta_y = \frac{1}{\Delta_v} \\ \Delta_x \le \frac{1}{L_u^F} \quad &\text{and} \quad \Delta_y \le \frac{1}{L_v^F} \\ M \ge \frac{L_x^f}{\Delta_x} \quad &\text{and} \quad N \ge \frac{L_y^f}{\Delta_y} \end{aligned} \tag{3.37}$$

This reveals that samples of $f(x,y)$ are directly related by the DFT operator to samples of $F(u,v)$. The first equation of (3.37) shows that the area spanned by the samples in the image plane ($M\Delta_x$ by $N\Delta_y$) is inversely proportional to the area of the resolution cell in the Fourier plane (Δ_u by Δ_v).

An efficient algorithm for calculating a DFT is the *fast Fourier transform (FFT) algorithm*, which is more efficient than other algorithms for computing a DFT because it minimizes the numbers of required complex additions and multiplications [Brigham, 1974, p. 151]. The form of the algorithm is simplest when both M and N are powers of 2 [Brigham, 1974, Chap. 11]. This form of the FFT algorithm is employed for all the examples presented in this thesis in which DFTs are evaluated.

3.4.2 Uniqueness of the Fourier phase problem

The Fourier phase problem, as posed in (3.17), has an infinity of different solutions. Obvious solutions are those images which share what has been called the same image-form as the correct solution. An image $g(x,y)$ is defined to have the same *image-form* as another image $f(x,y)$ if [Bates and McDonnell, 1989, Sec. 20]

$$g(x,y) = f(x-x_0, y-y_0)e^{j\psi_0} \quad \text{or} \quad g(x,y) = \tilde{f}(x-x_0, y-y_0)e^{j\psi_0} \tag{3.38}$$

where the arbitrary real constants x_0 and y_0 represent a translation of $f(x,y)$ and the arbitrary real number ψ_0 represents a constant phase term. The *conjugate reflection* of an image is denoted by a tilde and is defined by

$$\tilde{f}(x,y) = f^*(-x,-y) \tag{3.39}$$

Note that the form or appearance of an image is not altered if its phase is reversed, or if it is reflected through the origin, or if it is translated or multiplied by a complex constant. Table 3.3, which lists the relevant properties of Fourier transformation, confirms that all images with the same image-form also have the same Fourier transform amplitude, since

$$\begin{aligned} |G(x,y)| &= |F(u,v)e^{j2\pi(\psi_0+ux_0+vy_0)}| = |F(u,v)| \\ \text{or } |G(x,y)| &= |F^*(u,v)e^{j2\pi(\psi_0-ux_0-vy_0)}| = |F(u,v)| \end{aligned} \tag{3.40}$$

Because all of the image-forms of $f(x,y)$ are trivial solutions to (3.17), it is appropriate to redefine the Fourier phase problem as Bates and McDonnell [1989, Sec. 20]

$$\begin{gathered} \text{given } |F(u,v)|, \\ \text{retrieve the image-form of } f(x,y) \end{gathered} \tag{3.41}$$

A solution to the Fourier phase problem is said to be *unique* if there is no other image-form which has the required Fourier transform amplitude.

3.4.2.1 The Fourier transform amplitude

The Fourier phase problem requires knowledge of the amplitude $|F(u,v)|$ of the Fourier transform of an image $f(x,y)$. The amplitude $|F(u,v)|$ is directly related to the *autocorrelation* of $f(x,y)$, denoted by $ff(x,y)$ and defined in Table 3.3, via the *autocorrelation theorem* [Bracewell, 1978, Chap. 6; Bates and McDonnell, 1989, p. 25]:

$$ff(x,y) = \text{IFT}\Big\{|F(u,v)|^2\Big\} \tag{3.42}$$

Inspection of the autocorrelation integral (Table 3.3) reveals that, if $f(x,y)$ is exactly compact then so is $ff(x,y)$. Furthermore, the extents (Sec. 3.4.1.1) of the autocorrelation are given by [cf. Bates and McDonnell, 1989, Sec. 7]

$$L_x^{ff} = 2L_x^f \quad \text{and} \quad L_y^{ff} = 2L_y^f \tag{3.43}$$

Therefore, in order to satisfy the sampling theorem (Sec. 3.4.1.3) for $|F(u,v)|^2$, the sample spacings in the Fourier plane must satisfy (cf. (3.22))

$$\alpha_u\Delta_u = \frac{1}{L_x^f} \quad \text{and} \quad \alpha_v\Delta_v = \frac{1}{L_x^f} \tag{3.44}$$

where each of the sampling factors α_u and α_v is at least 2. This is known as *oversampling by a factor of at least two.* When using the DFT operator (Sec. 3.4.1.4), this ensures that the samples in the image plane span an area at least as large as the support of $ff(x,y)$.

At first sight it would seem that an infinity of image-forms can be solutions to the Fourier phase problem. Any phase function $\Psi(u,v)$ can be combined with $|F(u,v)|$ to generate an image

$$g(x,y) = \text{IFT}\Big\{|F(u,v)|e^{j\Psi(u,v)}\Big\} \tag{3.45}$$

However, assuming that $f(x,y)$ is exactly compact, its extents are derivable from $|F(u,v)|$ via (3.42) and (3.43). Only a small proportion of all possible $\Psi(u,v)$ are likely to result in an image $g(x,y)$ which has extents equal to those of $f(x,y)$. In fact, by utilizing the z-transform theory developed in the next section, it is shown in Section 3.4.2.4 that, almost always, any $g(x,y)$ generated by (3.45) which is of finite extent necessarily has the same image-form as $f(x,y)$.

3.4.2.2 The z-transform

As emphasized in Section 3.4.1, computer processing requires sampled images $f[m,n]$, which are related by the DFT operator to sampled Fourier transforms $F[p,q]$. The discrete equivalent of (3.42) is

$$ff[m,n] = \text{IDFT}\Big\{|F[p,q]|^2\Big\} \tag{3.46}$$

where the continuous Fourier transform amplitude is oversampled, by a factor of at least 2, to generate $|F[p,q]|$, and the *discrete autocorrelation* $ff[m,n]$ of $f[m,n]$ is defined in Table 3.4. It is therefore appropriate to repose the Fourier phase problem (3.41), for sampled images, as

$$\begin{gathered}\text{given } ff[m,n], \\ \text{retrieve the image-form of } f[m,n]\end{gathered} \tag{3.47}$$

In the following two sections, the solutions and uniqueness of the Fourier phase problem, as posed above, are discussed with the aid of the z-transform operation, which is introduced below.

For a sampled image $f[m,n]$, the two-dimensional, one sided, *z-transform* is here defined to be

$$\mathcal{F}(\zeta,\gamma) = \sum_{m=0}^{L_m^f}\sum_{n=0}^{L_n^f} f[m+m_{\min}, n+n_{\min}]\zeta^m\gamma^n \tag{3.48}$$

where ζ and γ are complex variables. With reference to Figure 3.8, $m_{\min}$ is the smallest value of m, and $n_{\min}$ is the smallest value of n, for which $f[m,n]$ is non-zero. Similarly, $(m_{\min}+L_m^f)$ and $(n_{\min}+L_n^f)$ are the largest values of m and n respectively, for which $f[m,n]$ is non-zero. The integers L_m^f and L_n^f are the discrete equivalent of the extents of the image (Sec. 3.4.1.1), while $m_{\min}$ and $n_{\min}$ determine the position of the sampled image in the x,y plane. The z-transform of a sampled image (e.g. $f[m,n]$) is denoted by the corresponding calligraphic uppercase letter (e.g. $\mathcal{F}(\zeta,\gamma)$).

The definition of the z-transform operation in (3.48) differs from many conventional definitions [e.g. Oppenheim and Schafer, 1975, Sec. 2.5] by the translation of $f[m,n]$ by $[-m_{\min}, -n_{\min}]$ and by the use of only positive powers of ζ and γ. As defined here, any z-transform is a *polynomial* in ζ and γ [Mostowski and Stark, 1964, Sec. IX-1-1]. A side effect is that $f[m,n]$ has the same z-transform as any translation $f[m-m_0, n-n_0]$ of $f[m,n]$, where m_0 and n_0 are arbitrary integers. Apart from this arbitrary translation, a sampled image can be uniquely determined from its z-transform [Oppenheim and Schafer, 1975, Sec. 2.2]. Relevant properties of the z-transform operation are listed in Table 3.4.

As an aside, it is interesting to note that the DFT and the z-transform of a sampled image are closely related. Comparing (3.33) with (3.48) reveals that

$$\begin{aligned} F[p,q] &= \Delta_x\Delta_y\zeta^{m_{\min}}\gamma^{n_{\min}}\mathcal{F}(\zeta,\gamma) \\ \text{where } \zeta &= e^{-j2\pi p/M} \\ \gamma &= e^{-j2\pi q/N} \end{aligned} \tag{3.49}$$

Various properties of polynomials $\mathcal{F}(\zeta,\gamma)$ are now stated. They are discussed in more detail by Mostowski and Stark [1964, Sec. IX-1]. A polynomial $\mathcal{F}(\zeta,\gamma)$ is said to be *irreducible* if there do not exist any two other, non-constant, polynomials $\mathcal{F}_1(\zeta,\gamma)$ and

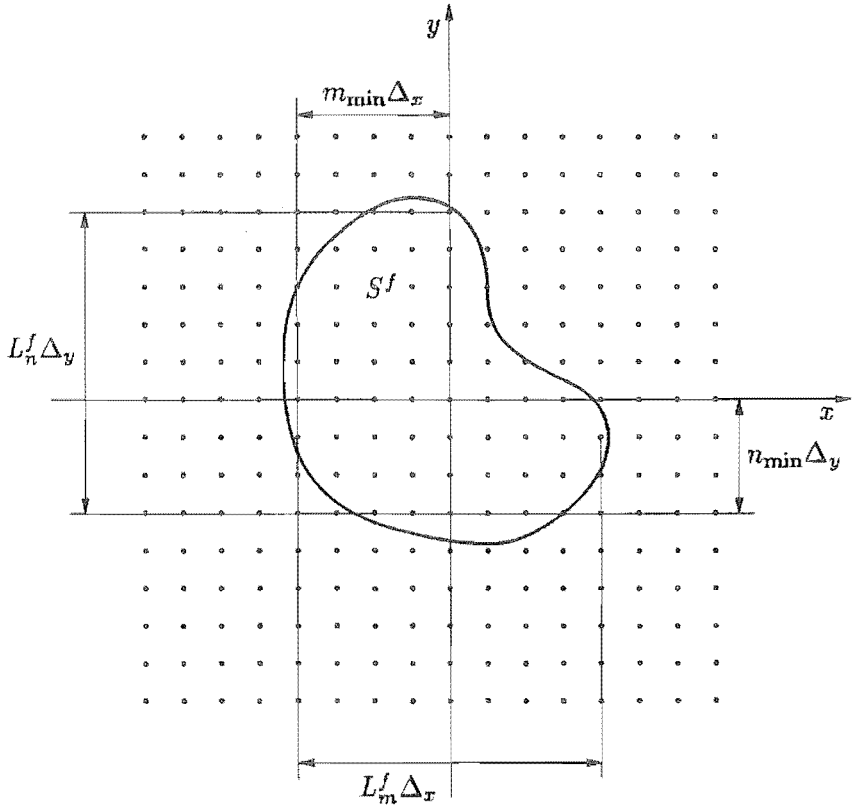


Figure 3.8 Definitions of $m_{\min}$, $n_{\min}$, L_m^F and L_n^F for a sampled image. The sampling grid (Sec. 3.4.1.2) and the support (Sec. 3.4.1.1) of the image are shown.

$\mathcal{F}_2(\zeta,\gamma)$ such that $\mathcal{F}(\zeta,\gamma) = \mathcal{F}_1(\zeta,\gamma)\,\mathcal{F}_2(\zeta,\gamma)$. A polynomial $\mathcal{F}(\zeta,\gamma)$ can be factored into a product of S non-constant polynomials $\mathcal{F}_s(\zeta,\gamma)$, each of which is irreducible:

$$\mathcal{F}(\zeta,\gamma) = \prod_{s=1}^{S} \mathcal{F}_s(\zeta,\gamma) \tag{3.50}$$

When $S = 1$, $\mathcal{F}(\zeta,\gamma)$ is itself irreducible.

When a polynomial can be factored in two different ways, such as

$$\begin{aligned} \mathcal{F}(\zeta,\gamma) &= \prod_{s=1}^{S} \mathcal{F}_s(\zeta,\gamma) \\ &= \prod_{t=1}^{T} \mathcal{G}_t(\zeta,\gamma) \end{aligned} \tag{3.51}$$

then $S = T$ necessarily, and the factors can always be ordered so that for all s, $\mathcal{F}_s(\zeta,\gamma) = c_s\mathcal{G}_s(\zeta,\gamma)$, where c_s is an arbitrary (complex) constant and $\prod_{s=1}^{S} c_s = 1$. Therefore, except for arbitrary constants, any z-transform (and therefore any sampled

Operation or function	Definition	Page
z-transformation	$\mathcal{F}(\zeta,\gamma) = \sum_{m=0}^{L_m'} \sum_{n=0}^{L_n'} f[m+m_{\min}, n+n_{\min}]\zeta^m\gamma^n$	[73]
convolution	$f[m,n] \odot g[m,n] = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} f[k,l]g[m-k,n-l]$	[61]
conjugate reflection	$\tilde{f}[m,n] = f^*[-m,-n]$	—
autocorrelation	$ff[m,n] = f[m,n] \odot \tilde{f}[m,n] = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} f[k,l]f^*[k-m,l-n]$	—

Property or operation	Sampled image	z-transform	Page
convolution	$f[m,n] \odot g[m,n]$	$\mathcal{F}(\zeta,\gamma)\mathcal{G}(\zeta,\gamma)$	[75]
conjugate reflection	$\tilde{f}[m,n]$	$\tilde{\mathcal{F}}(\zeta,\gamma) = \zeta^{L_m'}\gamma^{L_n'}\mathcal{F}^*\left(\frac{1}{\zeta^*},\frac{1}{\gamma^*}\right)$	[75]
autocorrelation	$ff[m,n]$	$\mathcal{F}(\zeta,\gamma)\tilde{\mathcal{F}}(\zeta,\gamma)$	—
translation	$f[m-k,n-l]$	$\mathcal{F}(\zeta,\gamma)$	—
scalar multiplication	$cf[m,n]$	$c\mathcal{F}(\zeta,\gamma)$	[60]

Table 3.4 Definitions and properties related to sampled images and their z-transforms (numbers in the last column refer to the relevant pages in Oppenheim and Schafer [1975]).

image) can be uniquely related to the set of irreducible polynomials which form its factorization.

This whole thesis rests on the remarkable fact, discussed in Section 3.4.2.4, that two-dimensional polynomials are almost always (i.e. except in special situations described in Sec. 3.4.2.4) irreducible. Section 3.4.2.3 demonstrates that, if the z-transform of $f[m,n]$ is irreducible, the Fourier phase problem as posed in (3.47) has one, and only one, solution. The implication is that solutions to the two-dimensional Fourier phase problem can generally be expected to be unique. As also pointed out in Section 3.4.2.4, this contrasts markedly with one-dimensional phase problems, solutions to which are almost always non-unique, because one-dimensional polynomials are never irreducible.

3.4.2.3 Solutions to the Fourier phase problem

Following Hayes [1982], it is now shown how all possible solutions to a Fourier phase problem (whether it possesses one or more than one solution) can be obtained. From (3.50) and the definitions in Table 3.4, the z-transform of the autocorrelation of $f[m,n]$ is

$$\begin{aligned} \mathcal{FF}(\zeta,\gamma) &= \mathcal{F}(\zeta,\gamma)\tilde{\mathcal{F}}(\zeta,\gamma) \\ &= \prod_{s=1}^{S} \mathcal{F}_s(\zeta,\gamma)\tilde{\mathcal{F}}_s(\zeta,\gamma) \end{aligned} \tag{3.52}$$

where the z-transforms of $f[m,n]$ and its conjugate reflection (cf. (3.39)) are related by (see Table 3.4)

$$\tilde{\mathcal{F}}(\zeta,\gamma) = \zeta^{L_m^I}\gamma^{L_n^I}\mathcal{F}^*\left(\frac{1}{\zeta^*},\frac{1}{\gamma^*}\right) \tag{3.53}$$

A similar relationship holds between $\mathcal{F}_s(\zeta,\gamma)$ and $\tilde{\mathcal{F}}_s(\zeta,\gamma)$, for each s.

The Fourier phase problem for sampled images (3.47) is equivalent to the problem of recovering $\mathcal{F}(\zeta,\gamma)$ from $\mathcal{FF}(\zeta,\gamma)$. It therefore reduces to the factorization a polynomial, with the factors related by (3.53) [Bruck and Sodin, 1979].

Consider a polynomial $\mathcal{G}(\zeta,\gamma)$ which is the z-transform of a solution to the Fourier phase problem. Expressing it as a product of irreducible factors, it must satisfy

$$\begin{aligned} \mathcal{FF}(\zeta,\gamma) &= \prod_{s=1}^{S} \mathcal{G}_s(\zeta,\gamma)\tilde{\mathcal{G}}_s(\zeta,\gamma) \\ \text{where } \mathcal{G}(\zeta,\gamma) &= \prod_{s=1}^{S} \mathcal{G}_s(\zeta,\gamma) \end{aligned} \tag{3.54}$$

Since $\mathcal{FF}(\zeta,\gamma)$ is uniquely defined by its irreducible factors (3.51), comparison of (3.54) with (3.52) gives, for each s,

$$\begin{aligned} \mathcal{G}_s(\zeta,\gamma)\tilde{\mathcal{G}}_s(\zeta,\gamma) &= c_s\mathcal{F}_s(\zeta,\gamma)c_s^*\tilde{\mathcal{F}}_s(\zeta,\gamma) \\ \text{and therefore } \mathcal{G}_s(\zeta,\gamma) &= c_s\mathcal{F}_s(\zeta,\gamma) \text{ or } c_s^*\tilde{\mathcal{F}}_s(\zeta,\gamma) \end{aligned} \tag{3.55}$$

where $\prod_{s=1}^{S} c_s c_s^* = 1$. Apart from the multiplier c_s, each factor of $\mathcal{G}(\zeta,\gamma)$ is one of two polynomials. Therefore, apart from a multiplier, there are at most 2^S different polynomials $\mathcal{G}(\zeta,\gamma)$ which are solutions of (3.54).

Each solution of (3.54) corresponds to a different image. However, if $\mathcal{G}(\zeta,\gamma)$ is a solution then so is $\tilde{\mathcal{G}}(\zeta,\gamma)$. Because both of these polynomials are z-transforms of the same image-form they constitute a single solution to the Fourier phase problem. This implies that there are at most 2^{S-1} solutions to the Fourier phase problem, where S is the number of irreducible factors of the z-transform of $f[m,n]$.

Instead of multiplying S irreducible polynomials to form the image's z-transform, one can equivalently convolve S sub-images to generate the complete image. Taking the inverse z-transform of (3.50) gives

$$f[m,n] = f_1[m,n] \odot f_2[m,n] \odot \ldots \odot f_S[m,n] \tag{3.56}$$

Each of the image-forms, which constitute the different solutions to the Fourier phase problem, is thus seen to be the convolution of one or more sub-images with the conjugate reflections of the remaining sub-images.

A study of the second equation of (3.54) and the second equation of (3.55) reveals that there can be can be fewer than 2^{S-1} solutions to the Fourier phase problem if any of the following conditions occur:

1. $\mathcal{F}(\zeta,\gamma)$ contains two factors $\mathcal{F}_s(\zeta,\gamma)$ and $\mathcal{F}_t(\zeta,\gamma)$ which are related by $\mathcal{F}_t(\zeta,\gamma) = c_s\mathcal{F}_s(\zeta,\gamma)$ where c_s is an arbitrary complex constant. This is equivalent to the image $f[m,n]$ having a repeated sub-image.

2. $\mathcal{F}(\zeta,\gamma)$ contains two factors $\mathcal{F}_s(\zeta,\gamma)$ and $\mathcal{F}_t(\zeta,\gamma)$ which are related by $\mathcal{F}_t(\zeta,\gamma) = c_s\tilde{\mathcal{F}}_s(\zeta,\gamma)$ where c_s is an arbitrary complex constant. This is equivalent to $f[m,n]$ having a sub-image proportional to the conjugate reflection of another sub-image.

3. $\mathcal{F}(\zeta,\gamma)$ contains any factor $\mathcal{F}_s(\zeta,\gamma)$ which possesses the property that $\mathcal{F}_s(\zeta,\gamma) = \tilde{\mathcal{F}}_s(\zeta,\gamma)$. This corresponds to $f[m,n]$ having a conjugate point symmetric sub-image.

An image or sub-image, say $f_1[m,n]$, is defined to be *conjugate point symmetric* if

$$f_1[m,n] = f_1^*[-m+m_0,-n+n_0] \tag{3.57}$$

where m_0 and n_0 are arbitrary integers. Note that a conjugate point symmetric image, centred about the origin, is equal to its conjugate reflection. Note also that the Fourier phase problem possesses a single solution either when $S = 1$ or when at least $(S-1)$ of the sub-images of $f[m,n]$ are conjugate point symmetric.

3.4.2.4 Images in one and two dimensions

A Fourier phase problem in one dimension almost always has several solutions [Bates and Mnyama, 1986, Sec. III-F; Taylor, 1981]. This is explained here in terms of z-transform theory. Although the theory presented in Sections 3.4.2.2 and 3.4.2.3 is developed for two-dimensional images, it also holds for one-dimensional images because one-dimensional images are a special case of two-dimensional images. The z-transform $\mathcal{F}(\zeta)$ of a one-dimensional sampled image $f[m]$ is a polynomial in only one complex variable. The smallest and largest values of m for which $f[m]$ is non-zero are, by definition, $m_{\min}$ and $(m_{\min}+L_m^f)$ respectively (Sec. 3.4.2.2). The *fundamental theorem of algebra* [Mostowski and Stark, 1964, Sec. VII-1] ensures that the polynomial $\mathcal{F}(\zeta)$ can always be factored into L_m^f factors. Each of these factors is of the form $c_s(\zeta-\zeta_s)$, where c_s is an arbitrary multiplier and ζ_s is called a *zero*. This means that a one-dimensional sampled image can always be expressed as a convolution of L_m^f sub-images, where each sub-image comprises two adjacent non-zero samples. It therefore follows from the reasoning presented in Section 3.4.2.3 that the Fourier phase problem for a one-dimensional image has up to $2^{L_m^f-1}$ solutions.

For a one-dimensional image $f[m]$ which comprises only two non-zero adjacent samples, the Fourier phase problem has only one solution because $L_m^f = 1$. However, most sampled images of interest contain more than two non-zero samples. The argument developed in the last paragraph of Section 3.4.2.3, when applied to one-dimensional z-transforms for which $L_m^f > 1$, indicates that the only way that a solution to the Fourier phase problem can be unique is for no more than one of the zeros ζ_s of $\mathcal{F}(\zeta)$ to satisfy $|\zeta_s| \neq 1$. Since this condition is not satisfied for a general one-dimensional z-transform, a solution to the one-dimensional Fourier phase problem is almost always non-unique. Mainly for this reason, one-dimensional images are not considered further in this thesis.

On the other hand, there is no fundamental theorem of algebra for polynomials of two variables. In fact, Hayes and McClellan [1982] show that almost all two-dimensional

polynomials are irreducible. This means that $S = 1$ for a general two-dimensional image. Therefore, a solution to the Fourier phase problem is almost always unique for two-dimensional images [Bates and Mnyama, 1986, Sec. III-F; Hayes and McClellan, 1982; [Bruck and Sodin, 1979]].

In special cases, however, $\mathcal{F}(\zeta,\gamma)$ can be reducible. This corresponds to $f[m,n]$ being a convolution of one or more sub-images. Consider an image $f[m,n]$ which is the convolution of two independent sub-images $f_1[m,n]$ and $f_2[m,n]$. Provided neither of the sub-images is conjugate point symmetric, there are two solutions to Fourier phase problem, which are the image-forms comprising scaled and translated versions of

$$\begin{aligned} g[m,n] &= f_1[m,n]f_2[m,n] \quad \text{or} \quad \tilde{f}_1[m,n]\tilde{f}_2[m,n] \\ \text{and} \quad g[m,n] &= f_1[m,n]\tilde{f}_2[m,n] \quad \text{or} \quad \tilde{f}_1[m,n]f_2[m,n] \end{aligned} \tag{3.58}$$

In practice, the sampled Fourier transform amplitude is usually obtained via a measurement. The accuracy to which the amplitude values are known is limited to the accuracy of the measurement. Let the measured Fourier transform amplitude be $A_{\mathrm{m}}[p,q] = (|F[p,q]| + N[p,q])$, where $N[p,q]$ represents a small amount of random noise. Sanz and Huang [1985] show that there is almost never a sampled image $g[m,n]$ whose discrete Fourier transform amplitude is equal to $A_{\mathrm{m}}[p,q]$. This is because $\mathrm{IDFT}\{(A_{\mathrm{m}}[p,q])^2\}$ is, in general, no longer an autocorrelation. An important implication of this is that the addition of noise to $|F[p,q]|$ does not introduce further solutions to the Fourier phase problem, and in that sense, the problem is stable.

3.4.3 Iterative Fourier transform algorithms

Any algorithm which attempts to solve the Fourier phase problem (see (3.41)) is called a *phase retrieval algorithm.* As pointed out by Dainty and Fienup [1987] the brute force algorithm, consisting of searching through all possible Fourier phase distributions, is impracticable. For example, if the Fourier transform is represented by 100 samples, and only 10 different values of phase are searched for each sample, a googol (10^{100}) phase distributions must be investigated — a high speed computer would be able to investigate only a negligible proportion of these during its lifetime. Therefore, very clever phase retrieval algorithms are required if they are to solve the Fourier phase problem in realistic times.

Phase retrieval algorithms can be either direct or iterative. Many different methods have been proposed and are the subject of a number of recent reviews [Fienup, 1982; Bates and Mnyama, 1986; Dainty and Fienup, 1987]. A large proportion of these methods are suitable for only special types of images, or in conjunction with specific measurement techniques [Dainty and Fienup, 1987]. The algorithms described in this section are suitable for any type of image, but can also incorporate any additional information about the image. Such information tends to increase the rate of convergence of the algorithms.

A direct method, based on the concepts discussed in Section 3.4.2.3, involves factorizing the z-transform $\mathcal{F}\mathcal{F}(\zeta,\gamma)$ into $\mathcal{F}(\zeta,\gamma)$ and $\tilde{\mathcal{F}}(\zeta,\gamma)$ (see (3.52)). An algorithm for achieving this, developed by Lane *et al.* [1987] and Lane and Bates [1987a], is now outlined in passing. Any polynomial $\mathcal{F}(\zeta,\gamma)$ is fully characterized by its zeros, here denoted by $Z^{\mathcal{F}}$, which are the set of points (ζ,γ) at which the polynomial vanishes. The zeros of irreducible two-dimensional polynomials form a single connected surface in (ζ,γ) space. Because $\mathcal{F}\mathcal{F}(\zeta,\gamma)$ is a product, $Z^{\mathcal{F}\mathcal{F}}$ is the union of $Z^{\mathcal{F}}$ and $Z^{\tilde{\mathcal{F}}}$. By utilizing the analytic properties of the polynomials, and their zeros, it is possible to

separate $Z^{\mathcal{F}}$ from $Z^{\tilde{\mathcal{F}}}$. From these $\mathcal{F}(\zeta,\gamma)$ and $\tilde{\mathcal{F}}(\zeta,\gamma)$ can then be calculated. In the form outlined above, this method fails when, as is inevitable in practice, the data are noisy. This failure occurs because the polynomial corresponding to $\mathcal{F}\mathcal{F}(\zeta,\gamma)$ is irreducible, as intimated in the final paragraph of Section 3.4.2.4. Lane [1987] points out that iterative methods, such as the one about to be discussed, tend to be more robust than direct methods.

The algorithms described in the next two sections are special forms of the *basic iterative Fourier transform algorithm* which is depicted in Figure 3.9. This algorithm iterates between the image and Fourier planes, applying constraints in each plane. The constraints represent available, but incomplete, information about $f(x,y)$ and $F(u,v)$, denoted by $[f(x,y)]$ and $[F(u,v)]$ respectively. In order to solve the Fourier phase problem, this information must contain at least $|F(u,v)|$ and the extents of $f(x,y)$, the latter of which can be deduced from the former through (3.42) and (3.43). At the i^{th} iteration, the i^{th} image, $g_i(x,y)$, is Fourier transformed, to obtain $G_i(u,v)$. This is then constrained by $[F(u,v)]$ to produce $G_i'(u,v)$. An inverse Fourier transformation produces $g_i'(x,y)$, which is constrained by $[f(x,y)]$ to produce the $(i+1)^{\text{th}}$ image. The application of constraints can, in general, be any operation which utilizes the information contained in $[f(x,y)]$ or $[F(u,v)]$. The constraints applied in the image and Fourier planes are called the *image constraints* and *Fourier constraints* respectively. The degree of convergence to a solution is indicated by the difference between $|G_i(u,v)|$ and $|F(u,v)|$.

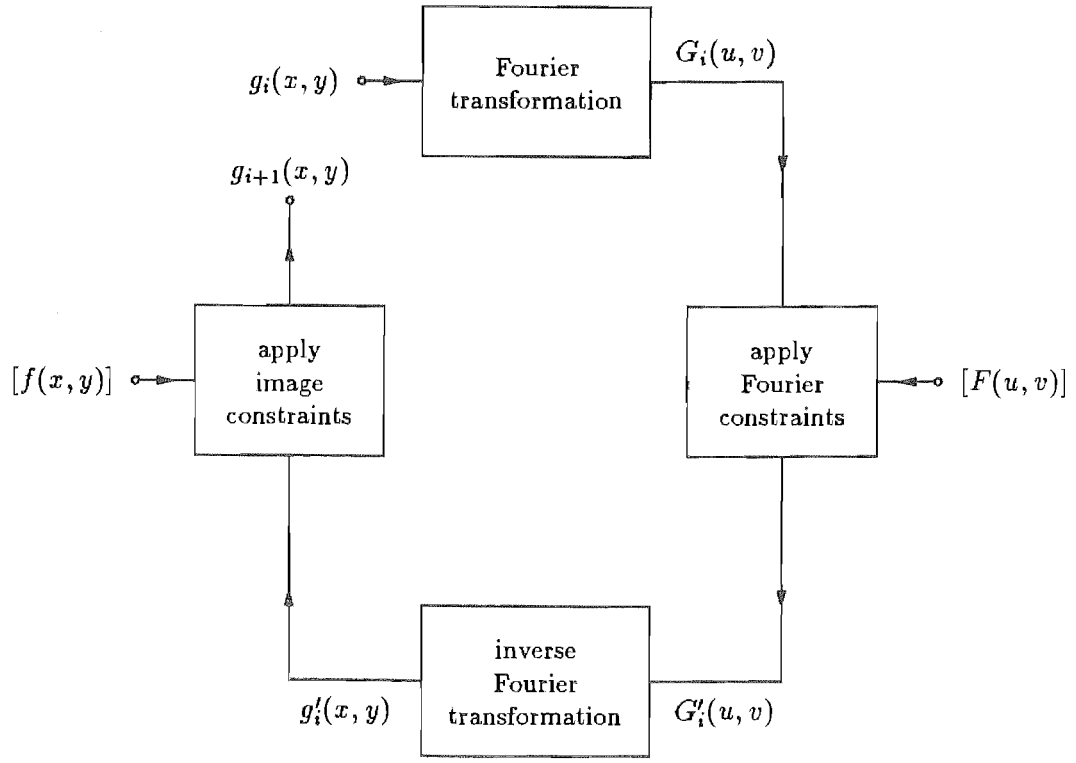


Figure 3.9 One iteration of the basic iterative Fourier transform algorithm. The boxes represent operations and the labelled arrows indicate the inputs to and the outputs from these operations.

The Gerchberg-Saxton algorithm, described in Section 3.4.3.1, can be invoked when the amplitude distribution of the image is available. It forms the starting point for the development of the modified Gerchberg-Saxton algorithm described in Chapter 4. A variant of the Gerchberg-Saxton algorithm is described in Section 3.4.3.2. In Section 3.4.3.3 various algorithms developed by Fienup are described. They were initially applied to non-negative, real images, but are now known to be applicable to complex-valued images, which are here called *complex images.*

The Fourier phase problem is not the only problem to which the basic iterative Fourier transform algorithm can applied. The algorithm can also be employed to extrapolate bandlimited data (Sec. 3.5.6), synthesize aperture distributions to produce a desired radiation pattern [Fienup, 1980] and retrieve $f(x,y)$ when phase$\{F(u,v)\}$ is given [e.g. Lane and Bates, 1987b]. The latter problem can be considered the complement of the Fourier phase problem. The algorithm can also be extended to blindly deconvolve a noisy convolution of complex images [e.g. Davey *et al.*, 1989]. Other kinds of phase retrieval algorithm, which are applicable to the radio engineering phase problem, are discussed in Section 3.5.

In the following sections, and throughout this thesis, the notations $f(x,y)$ and $F(u,v)$ imply continuous distributions over the x,y and u,v planes respectively. However, it must be remembered that, as pointed out in Section 3.4.1.2, images of interest in practical applications are always sampled. It is assumed that the conditions introduced in (3.37) are met, so that the Fourier transform of a continuous image is equivalent to the DFT of the corresponding sampled image. Similarly, integrals are computed as summations. When no limits are specified on the integrals, they are assumed to be evaluated over the region spanned by the grid of sample points.

3.4.3.1 The Gerchberg-Saxton algorithm

The *Gerchberg-Saxton algorithm* was originally developed for electron microscopy [Gerchberg and Saxton, 1972; Saxton, 1978, Secs. 5.3 and 6.3]. However, it can be applied to the Fourier phase problem wherever it arises, provided that the amplitudes of the image as well as its Fourier transform are available. In terms of the basic iterative Fourier transform algorithm, the available information is $[f(x,y)] = |f(x,y)|$ and $[F(u,v)] = |F(u,v)|$. One iteration of the Gerchberg-Saxton algorithm is described by the following equations:

$$\begin{aligned} G_i(u,v) &= \mathrm{FT}\{g_i(x,y)\} \\ G_i'(u,v) &= |F(u,v)|e^{j\,\mathrm{phase}\{G_i(u,v)\}} \\ g_i'(x,y) &= \mathrm{IFT}\{G_i'(u,v)\} \\ g_{i+1}(x,y) &= |f(x,y)|e^{j\,\mathrm{phase}\{g_i'(x,y)\}} \end{aligned} \tag{3.59}$$

Notice that the second and fourth equations of (3.59) ensure that $G_i'(u,v)$ and $g_{i+1}(x,y)$ respectively are constrained to possess the available amplitudes.

Gerchberg and Saxton [1972] start their algorithm with $g_1(x,y)$, whose amplitude equals $|f(x,y)|$ and whose phase is randomly distributed between $-\pi$ and π. The degree of convergence of the algorithm is indicated by the *Fourier error* $\mathcal{E}_i^{\mathrm{F}}$, which can

be calculated at each iteration, and is defined by

$$\mathcal{E}_i^{\mathrm{F}} = \left[\frac{\iint [|G_i(u,v)| - |F(u,v)|]^2 \, du\, dv}{\iint |F(u,v)|^2 \, du\, dv} \right]^{1/2} \tag{3.60}$$

Therefore, $\mathcal{E}_i^{\mathrm{F}}$ is a measure of how close $|G_i(u,v)|$ is to $|F(u,v)|$. Obviously, if $\mathcal{E}_i^{\mathrm{F}} = 0$ then $g_i(x,y)$ is an exact solution to the Fourier phase problem. However, the available information $[f(x,y)]$ and $[F(u,v)]$ invariably contains uncertainties, making an exact solution impossible (see last paragraph in Sec. 3.4.2.4). So the algorithm is deemed to have *converged* when $\mathcal{E}_i^{\mathrm{F}}$ falls below a preset level, which is chosen to reflect the noise level in the data.

The convergence properties of the Gerchberg-Saxton algorithm are discussed by Gerchberg and Saxton [1972] and are demonstrated here with the aid of Figure 3.10. Note from Figure 3.10(a) that the distance between $g_i'(x,y)$ and $g_{i+1}(x,y)$ can never be greater than the distance between $g_i'(x,y)$ and $g_i(x,y)$. This is true for all possible values of $g_i(x,y)$, $g_i'(x,y)$ and $g_{i+1}(x,y)$, provided that the latter two are related by the last equation of (3.59). Application of this inequality to all points in the x,y plane makes the following inequality hold:

$$\iint |g_i'(x,y) - g_{i+1}(x,y)|^2 \, dx\, dy \leq \iint |g_i'(x,y) - g_i(x,y)|^2 \, dx\, dy \tag{3.61}$$

From the energy conservation theorem for Fourier transforms [Bates and McDonnell, 1989, p. 24], each side of (3.61) can be Fourier transformed to give

$$\iint |G_i'(u,v) - G_{i+1}(u,v)|^2 \, du\, dv \leq \iint |G_i'(u,v) - G_i(u,v)|^2 \, du\, dv \tag{3.62}$$

Note that, from Figure 3.10(b), the integrand on the right side of (3.62) is equal to $(|G_i(u,v)| - |F(u,v)|)^{1/2}$. Applying to Figure 3.10(b) the same reasoning as is applied above to Figure 3.10(a), the following inequality is established:

$$\iint (|G_{i+1}(u,v)| - |F(u,v)|)^2 \, du\, dv \leq \iint |G_{i+1}(u,v) - G_i'(u,v)|^2 \, du\, dv \tag{3.63}$$

Substituting (3.62) into (3.63), and invoking (3.60), gives

$$\mathcal{E}_{i+1}^{\mathrm{F}} \leq \mathcal{E}_i^{\mathrm{F}} \tag{3.64}$$

which demonstrates that the Gerchberg-Saxton algorithm can never diverge. However, it is possible for the algorithm to converge extremely slowly, in which case the algorithm is said to have *stagnated.* The above reasoning shows that stagnation occurs whenever, but only when, $g_{i+1}(x,y)$ is almost equal to $g_i(x,y)$ at all points (x,y).

If the Gerchberg-Saxton algorithm converges, to within a preset level, after I iterations, $g_I(x,y)$ is taken to be an estimate of the image-form (defined in Sec. 3.4.2) of $f(x,y)$. Because, from the fourth equation of (3.59), $|g_I(x,y)| = |f(x,y)|$, the constants x_0 and y_0 in (3.38) must be zero. The image $g_I(x,y)$ can be an estimate of either $f(x,y)e^{j\psi_0}$ or of $\tilde{f}(x,y)e^{j\psi_0}$, where ψ_0 is an arbitrary real value, only if $|f(x,y)|$ is point symmetric. An image, say $q(x,y)$, is defined to be *point symmetric (images)* if

$$q(x,y) = q(-x,-y) \tag{3.65}$$

over the whole x,y plane. If $|f(x,y)|$ is not point symmetric, $g_I(x,y)$ is taken to be an estimate of $f(x,y)$ to within a constant phase term.

An example employing the Gerchberg-Saxton algorithm is given in Section 4.4.1.

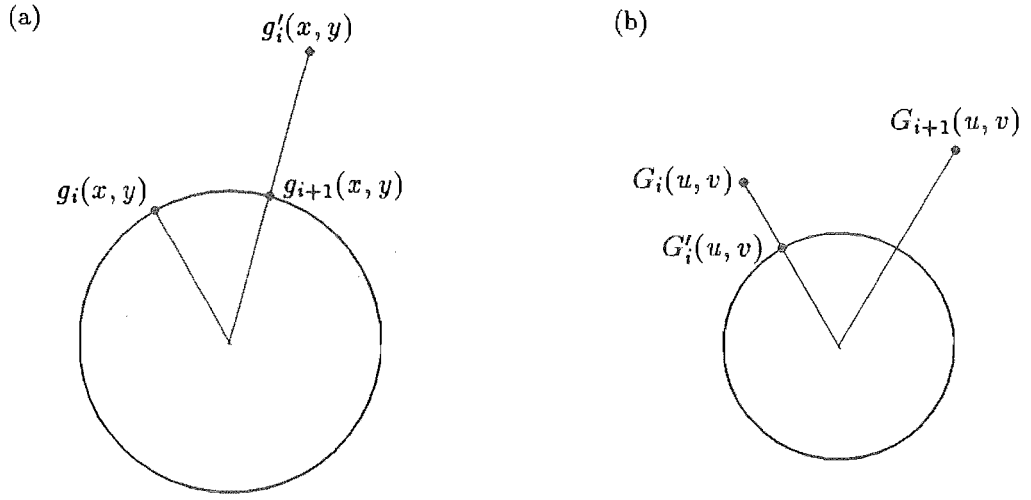


Figure 3.10 Action of the Gerchberg-Saxton algorithm at typical points in (a) the x, y plane and (b) the u, v plane. The circles in (a) and (b) are the loci of complex numbers having amplitudes of $|f(x,y)|$ and $|F(u,v)|$ respectively.

3.4.3.2 A variant of the Gerchberg-Saxton algorithm

A disadvantage of the original Gerchberg-Saxton algorithm (Sec. 3.4.3.1) is that in practice it often stagnates far from a solution to the Fourier phase problem. This can be explained by recasting the Fourier phase problem as a minimization problem: find an image $g(x,y)$ that minimizes $\mathcal{E}^{\mathrm{F}}$, subject to the constraint $|g(x,y)| = |f(x,y)|$. Stagnation corresponds to $\mathcal{E}_i^{\mathrm{F}}$ being close to a local minimum. Since, in the Gerchberg-Saxton algorithm, $\mathcal{E}_i^{\mathrm{F}}$ cannot increase, the algorithm can never progress out of the local 'well' and on to the global minimum.

Gerchberg [1986] has suggested a variant of the original Gerchberg-Saxton algorithm in which $\mathcal{E}_i^{\mathrm{F}}$ is allowed to increase between iterations, but is bound to lie below a parameter which itself decreases as the number of iterations increases. The variation from the original Gerchberg-Saxton algorithm (Sec. 3.4.3.1) is in the way that the constraints are applied in the image and Fourier planes. The second and fourth equations of (3.59) are replaced by

$$\begin{aligned} G_i'(u,v) &= |F(u,v)| e^{j\,\mathrm{phase}\{G_i(u,v)\}} e^{j\gamma_i^{\mathrm{F}}(u,v)[\mathrm{phase}\{G_{i-1}'(u,v)\} - \mathrm{phase}\{G_i(u,v)\}]} \\ g_{i+1}(x,y) &= |f(x,y)| e^{j\,\mathrm{phase}\{g_i'(x,y)\}} e^{j\gamma_i^{\mathrm{I}}(x,y)[\mathrm{phase}\{g_i(x,y)\} - \mathrm{phase}\{g_i'(x,y)\}]} \end{aligned} \tag{3.66}$$

where $\gamma_i^{\mathrm{I}}(x,y)$ and $\gamma_i^{\mathrm{F}}(u,v)$ are real, are constrained to lie between -1 and 1, and can be chosen independently from iteration to iteration and from point to point in the x, y and u, v planes respectively. Note that, when $\gamma_i^{\mathrm{I}}(x,y) = \gamma_i^{\mathrm{F}}(u,v) = 0$, the algorithm reduces to the original Gerchberg-Saxton algorithm.

The convergence characteristics of the above variant of the Gerchberg-Saxton algorithm can be studied with the aid of Figure 3.11. This figure also helps to visualize the computations implied by (3.66). For example, the second equation of (3.66) corresponds to Figure 3.11(a). The value of $g_{i+1}(x,y)$ can lie anywhere on the arc, indicated in Figure 3.11(a) by a thick curve, depending on the value of $\gamma_i^{\mathrm{I}}(x,y)$. The limits of the

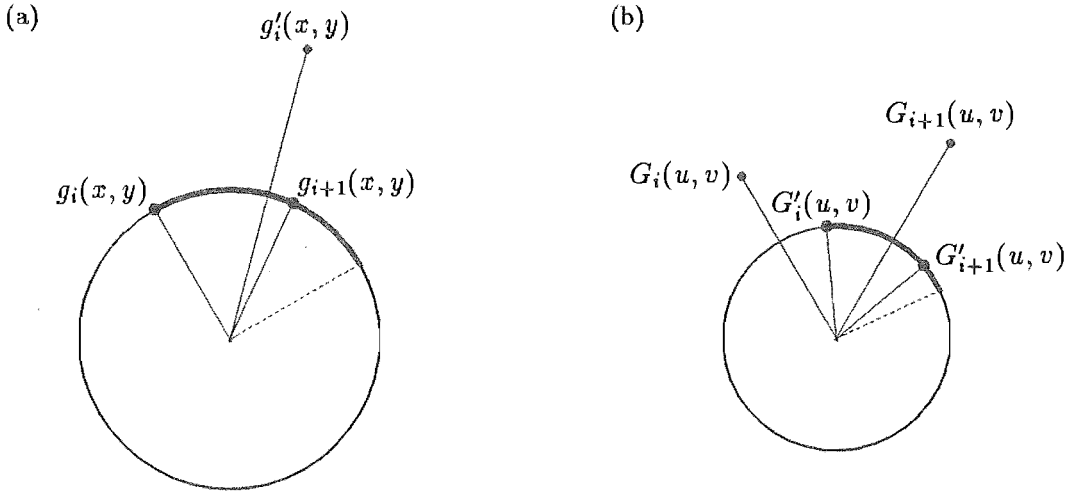


Figure 3.11 Action of the variant of the Gerchberg-Saxton algorithm, introduced in Section 3.4.3.2, at typical points in (a) the x, y plane and (b) the u, v plane. The circles in (a) and (b) are the loci of complex numbers with amplitudes of $|f(x,y)|$ and $|F(u,v)|$ respectively. The arcs indicated by thick curves in (a) and (b) indicate possible values of g_{i+1} and G_{i+1}' respectively.

arc are such that the distance between $g_i'(x,y)$ and $g_{i+1}(x,y)$ is never greater than the distance between $g_i'(x,y)$ and $g_i(x,y)$. Therefore, (3.61) and (3.62) necessarily hold. Similarly, from Figure 3.11(b), (3.63) must also hold, as must

$$\iint |G_{i+1}(u,v) - G_{i+1}'(u,v)|^2 \, du\, dv \leq \iint |G_{i+1}(u,v) - G_i'(u,v)|^2 \, du\, dv \qquad (3.67)$$

It is now convenient to introduce the normalized rms Fourier correction C_i^{F} which is defined to be

$$C_i^{\mathrm{F}} = \left[\frac{\iint |G_i(u,v) - G_i'(u,v)|^2 \, du\, dv}{\iint |F(u,v)|^2 \, du\, dv} \right]^{1/2} \qquad (3.68)$$

It follows from equations (3.60) to (3.63), (3.67) and (3.68) that

$$C_{i+1}^{\mathrm{F}} \leq C_i^{\mathrm{F}} \quad \text{and} \quad \mathcal{E}_i^{\mathrm{F}} \leq C_i^{\mathrm{F}} \qquad (3.69)$$

The rate at which C_i^{F} decreases depends in part on $\gamma_i^{\mathrm{I}}(x,y)$ and $\gamma_i^{\mathrm{F}}(u,v)$. If $|\gamma_i^{\mathrm{I}}(x,y)|$ and $|\gamma_i^{\mathrm{F}}(u,v)|$ are always equal to 1, C_i^{F} stays constant. Because the Fourier error $\mathcal{E}_i^{\mathrm{F}}$ can assume any positive value less than C_i^{F}, it need not necessarily be trapped near local minima.

An example of the above variant of the Gerchberg-Saxton algorithm is presented in Section 4.4.2.

3.4.3.3 Fienup's algorithms

The algorithms developed by Fienup [1978] are adaptations of the Gerchberg-Saxton algorithm (Sec. 3.4.3.1) which are suitable for application to *positive images* (i.e. images whose sample values are real and non-negative [Bates and McDonnell, 1989, p. 27]).

In these situations, $[F(u,v)] = |F(u,v)|$, while $[f(x,y)]$ contains information about the support and the positive nature of $f(x,y)$. If the support S^f of $f(x,y)$ is not available, it is often taken to be the region enclosed by a rectangle whose sides are parallel to the Cartesian axes and are equal in length to the extents of $f(x,y)$. These extents can be calculated from $|F(u,v)|$ through (3.42) and (3.43). The support invoked for the algorithms is here denoted by $S^{[f]}$, to distinguish it from S^f.

The most obvious adaptation of the Gerchberg-Saxton algorithm is called the *error reduction algorithm* [Fienup, 1982]. It is described by (3.59) with the last equation replaced by

$$g_{i+1}(x,y) = \begin{cases} g_i'(x,y) & \text{for } (x,y) \in \Upsilon_i \\ 0 & \text{elsewhere} \end{cases} \tag{3.70}$$

where Υ_i is the set of points (x,y) at which $g_i'(x,y)$ satisfies the image constraints. When the image is known to be positive, Υ_i consists of all points $(x,y) \in S^{[f]}$ at which $g_i'(x,y)$ is positive. As for the Gerchberg-Saxton algorithm, the degree of convergence is monitored by $\mathcal{E}_i^{\mathrm{F}}$, which is defined in (3.60). Another indication of the degree of convergence, which is utilized later in this section, is the *image error* $\mathcal{E}_i^{\mathrm{I}}$ defined by

$$\mathcal{E}_i^{\mathrm{I}} = \left[\frac{\iint_{(x,y)\notin\Upsilon_i} |g_i'(x,y)|^2 \, dx\, dy}{\iint |f(x,y)|^2 \, dx\, dy} \right]^{1/2} \tag{3.71}$$

Thus $\mathcal{E}_i^{\mathrm{I}}$ is a measure of how much $g_i'(x,y)$ violates the image constraints implied by $[f(x,y)]$. The denominator in (3.71) can be calculated from $|F(u,v)|$ by employing the energy conservation theorem for Fourier transforms [Bates and McDonnell, 1989, p. 24].

Employing a similar approach to that presented in Section 3.4.3.1, Fienup [1982] has proved that the error reduction algorithm, like the Gerchberg-Saxton algorithm, can never diverge. This property has given the error reduction algorithm its name, because the errors $\mathcal{E}_i^{\mathrm{F}}$ and $\mathcal{E}_i^{\mathrm{I}}$ always reduce (even if infinitesimally on occasion) between iterations. As intimated in Section 3.4.3.2, this property implies that the error reduction algorithm is prone to stagnation. In practice, the error reduction algorithm tends to converge more slowly than the Gerchberg-Saxton algorithm [Fienup, 1982].

In order to overcome stagnation, Fienup has investigated alternative methods of applying the constraints in the image plane. In all of these methods, the first three equations of (3.59) are regarded as describing a nonlinear process, with an input $g_i(x,y)$ and an output $g_i'(x,y)$. For this particular process, small changes in the input are expected to produce similar changes (both in amplitude and phase) in the output [Fienup, 1980, Appendix]. Rather than forcing $g_{i+1}(x,y)$ to satisfy the image constraints, it is chosen in such a way as to drive $g_{i+1}'(x,y)$ towards satisfying the constraints. Therefore, in the *input-output algorithm* [Fienup, 1978], the last equation of (3.59) is replaced by

$$g_{i+1}(x,y) = \begin{cases} g_i(x,y) & \text{for } (x,y) \in \Upsilon_i \\ g_i(x,y) - \beta g_i'(x,y) & \text{elsewhere} \end{cases} \tag{3.72}$$

where β is a real constant, called the *feedback parameter*, which is usually chosen to lie between 0 and 1. If, after I iterations, the algorithm has converged to within a preset level, $g_I'(x,y)$ is taken to be the estimate, generated by the algorithm, of the image-form of $f(x,y)$. This is appropriate because, not only does $g_I'(x,y)$ approximately meet the image constraints, its Fourier transform $G_I'(u,v)$ also exactly meets

the Fourier constraints by virtue of the second equation of (3.59). Even if this algorithm converges exactly, $|G_I(u,v)|$ need not resemble $|F(u,v)|$ — all that is required is for phase$\{G_I(u,v)\}$ to equal phase$\{F(u,v)\}$ to within a constant. This makes $\mathcal{E}_i^{\mathrm{F}}$ meaningless as an indication of the degree of convergence of this algorithm, so instead $\mathcal{E}_i^{\mathrm{I}}$ must always be utilized.

The *hybrid input-output algorithm* [Fienup, 1982] is a combination of the error reduction and input-output algorithms. It is described by (3.59) with the last equation replaced by (cf. (3.70) and (3.72))

$$g_{i+1}(x,y) = \begin{cases} g_i'(x,y) & \text{for } (x,y) \in \Upsilon_i \\ g_i(x,y) - \beta g_i'(x,y) & \text{elsewhere} \end{cases} \tag{3.73}$$

For the reasons given in the previous paragraph, the degree of convergence is indicated by $\mathcal{E}_i^{\mathrm{I}}$. Also, if the algorithm converges to within a preset level after I iterations, $g_I'(x,y)$ is taken to be the generated estimate of the image-form of $f(x,y)$. Although no supporting theoretical analysis has been developed for the convergence properties of the hybrid input-output algorithm, computational experience reveals that it is the most successful of the phase retrieval algorithms developed by Fienup [1982].

The Fourier phase problem for positive images can almost always be solved, in a straightforward manner, by employing the hybrid input-output algorithm [Bates and Mnyama, 1986]. In practice, the hybrid input-output algorithm can stagnate, although very rarely by comparison with the error reduction algorithm. Fienup and Wackerman [1987] have studied the causes of stagnation in the hybrid input-output algorithm and have developed methods for avoiding them.

An important advantage of iterative algorithms, such as the hybrid input-output algorithm, is that they are robust in the presence of noise: the rms error in the final estimate of the image-form is roughly equal to the square root of the rms error of the Fourier amplitude estimate [Feldkamp and Fienup, 1980]. However, when the data are appreciably contaminated, the estimate of the image-form, generated by the hybrid input-output algorithm, does not steadily improve as the number of iterations increases. Furthermore, $\mathcal{E}_i^{\mathrm{I}}$ tends to fluctuate erratically. McCallum and Bates [1989] have observed, nevertheless, that the $g_i'(x,y)$ corresponding to the several locally minimum values of $\mathcal{E}_i^{\mathrm{I}}$ exhibit similarities to the image-form of $f(x,y)$. They have devised a technique for appropriately averaging these $g_i'(x,y)$ to provide an estimate of the image which is more faithful than any estimate generated directly by the hybrid input-output algorithm.

It has recently been discovered that the algorithms described in this section can also be successfully applied to the Fourier phase problem for complex images. In this case, Υ_i is replaced by $S^{[f]}$ in (3.70), (3.72) and (3.73). It was initially thought that complex images could only be recovered if they possessed specialized supports [Fienup, 1987]. However, Lane [1987] has demonstrated the retrieval of complex images with other, more common, supports. Application of the error reduction and hybrid input-output algorithms to a complex image is now demonstrated with the aid of Figures 3.12 to 3.15. The sampled images and sampled Fourier transforms manipulated by the algorithms each contain 128 by 128 samples. The support $S^{[f]}$ is a centred rectangle spanning 62 by 58 samples which just encloses S^f. The ratio of the support extent to the extent of the region spanned by the image plane samples, in each of the x and y directions, ensures that $|F(u,v)|$ is oversampled by a factor greater than two (which is required for reasons given in Sec. 3.4.2.1). Only the samples within $S^{[f]}$ are displayed for each of the images

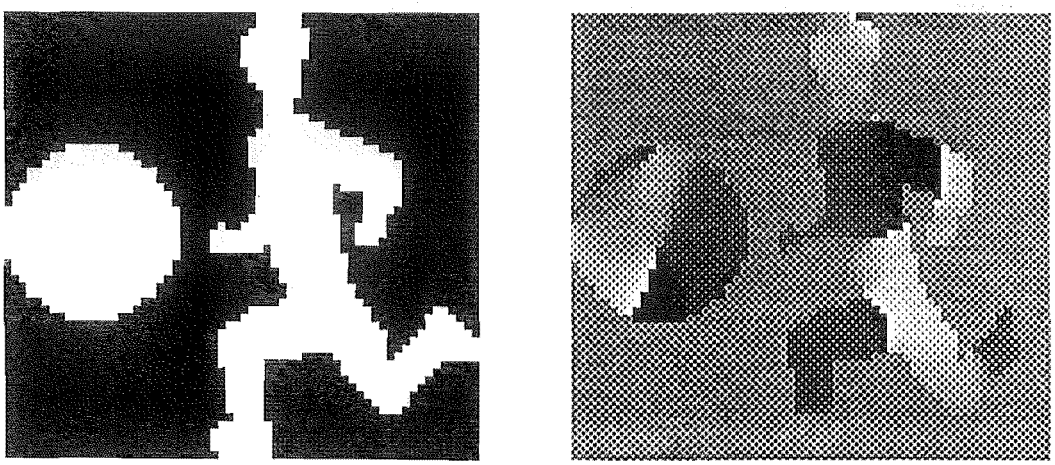
depicted in Figures 3.12 to 3.14. The true image $f(x,y)$ and the starting image $g_1(x,y)$, employed by both algorithms, are shown in Figure 3.12. The Fourier information $[F(u,v)]$ utilized by both algorithms is the true Fourier transform amplitude $|F(u,v)|$. Figures 3.13 and 3.14 depict $g_i(x,y)$, at various iterations, for the error reduction and hybrid input-output algorithms respectively. Comparison of Figure 3.13(c) with Figure 3.14(c) shows that the hybrid input-output algorithm produces a faithful replica of $f(x,y)$, while in the same number of iterations, the error reduction algorithm does not. A graph of the variation of $\mathcal{E}_i^{\mathrm{I}}$ versus the number of iterations, for both algorithms, is plotted in Figure 3.15. It is clearly seen that the error reduction algorithm stagnates with a relatively large value of $\mathcal{E}_i^{\mathrm{I}}$. On the other hand, although $\mathcal{E}_i^{\mathrm{I}}$ fluctuates erratically for the hybrid input-output algorithm, it has an overall converging trend. Note that the solution generated by the hybrid input-output algorithm (Fig. 3.14(c)) is an estimate of the conjugate reflection of $f(x,y)$ multiplied by a constant phase term. Because this solution has the same image-form as $f(x,y)$, it is a satisfactory solution to the Fourier phase problem (see Sec. 3.4.2).

In general, retrieval of complex images requires more iterations than for positive images. Both Lane [1987] and Fienup [1987] note that the image error $\mathcal{E}_i^{\mathrm{I}}$ has a poor correlation with image quality (e.g. compare Fig. 3.14(b) with Fig. 3.13(b) which correspond to values of $\mathcal{E}_i^{\mathrm{I}}$ that are approximately equal). Observing that $\mathcal{E}_i^{\mathrm{I}}$ is more correlated with image quality for the error reduction algorithm than for the hybrid input-output algorithm, Fienup [1982] developed a cycling technique for applying his algorithms: one cycle consists of a number, say 40, of hybrid input-output iterations, followed by say 10 error reduction iterations. The cycle is repeated until the algorithm either converges to a preset level or stagnates. This has the advantage that, at the end of each cycle, $\mathcal{E}_i^{\mathrm{I}}$ best indicates the faithfulness of the corresponding image estimate. However, Lane [1987] claims that the rate of convergence is usually faster when only the hybrid input-output algorithm is invoked.

Lane [1987] also demonstrates that the hybrid input-output algorithm can successfully retrieve a complex image when $S^{[f]}$ is larger than S^f, at the expense of an increased number of iterations. Cederquist *et al.* [1988] have successfully demonstrated phase retrieval from experimentally obtained optical data, when the image, which was complex, had a specialized support and the cycling technique was employed.

Because of the large number of iterations required to retrieve a complex image, any practical means of reducing the computation time is worth developing. McCallum and Bates [1989] have observed that, at any given iteration, the hybrid input-output algorithm tends to have retrieved phase$\{F(u,v)\}$ over a region of the u,v plane which is roughly centred on the origin. This region is at first small, but grows with increasing numbers of iterations. This suggests that the algorithm could be initially applied merely to a restricted number of central samples of the Fourier amplitude and a correspondingly coarsely sampled image. Because it would then be operating on a small number of samples, the algorithm should proceed relatively quickly. After $\mathcal{E}_i^{\mathrm{I}}$ is sufficiently small, more Fourier amplitude samples can be included, and the algorithm rerun, starting with the centre Fourier phases being those obtained from the previous run. Then even more samples can be included, with the procedure being repeated until all the samples are operated upon. McCallum and Bates [1989] present an example of this procedure for a complex image, demonstrating that about half the previously needed computer time can be saved.

(a)



(b)

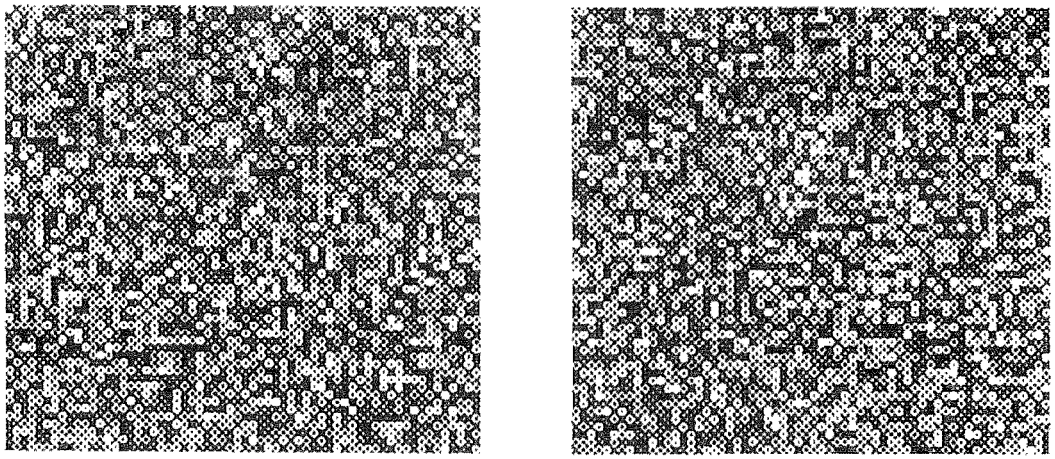
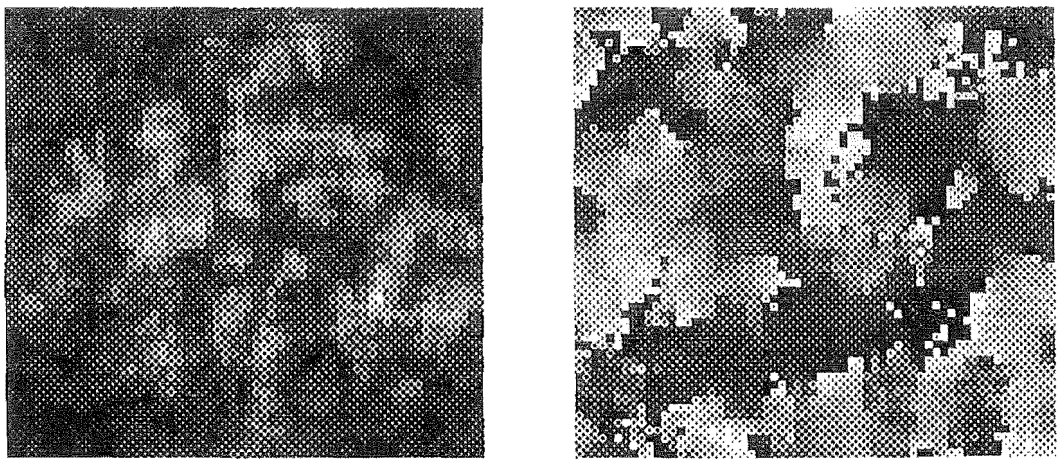
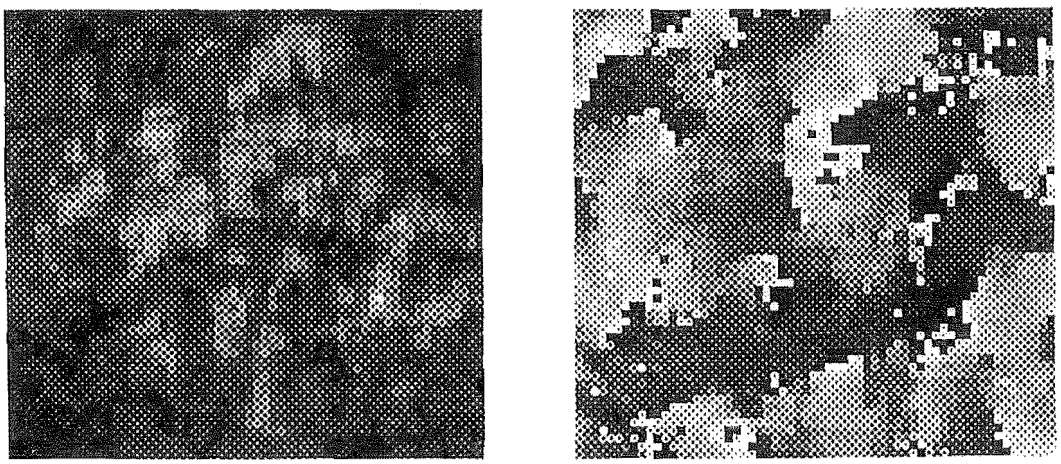


Figure 3.12 Images used in examples of Fienup's algorithms: (a) complex image $f(x, y)$; (b) random starting image $g_1(x, y)$. The figures on the left show the amplitude ranging from 0 (black) to maximum (white) and those on the right show phases ranging from $-\pi$ (black) to π (white).

(a)



(b)



(c)

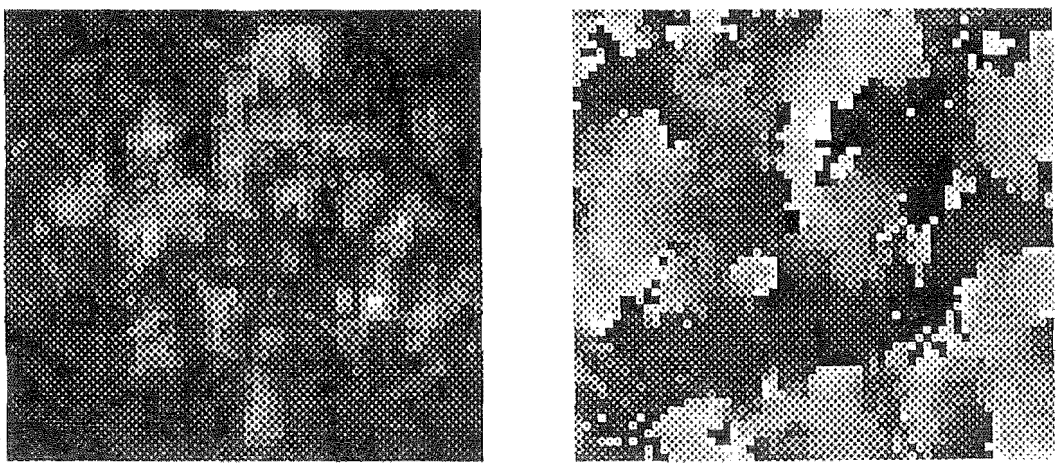
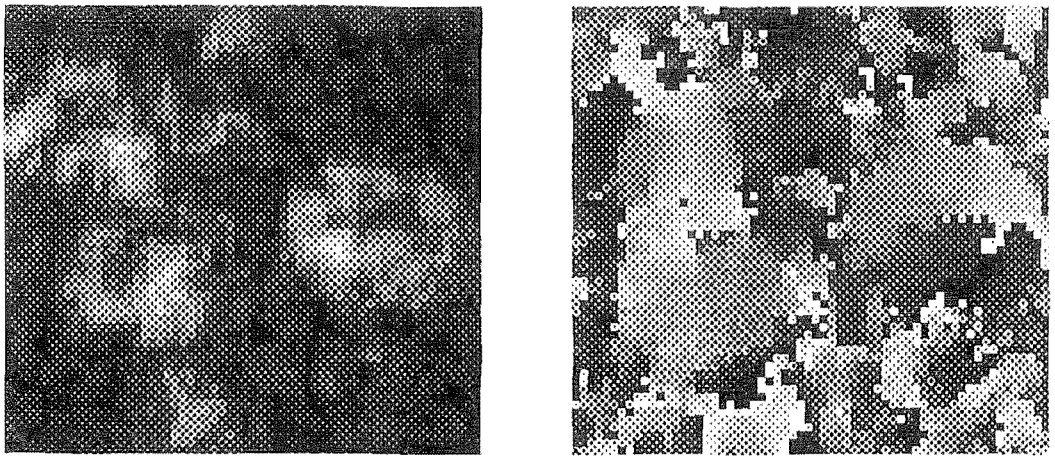
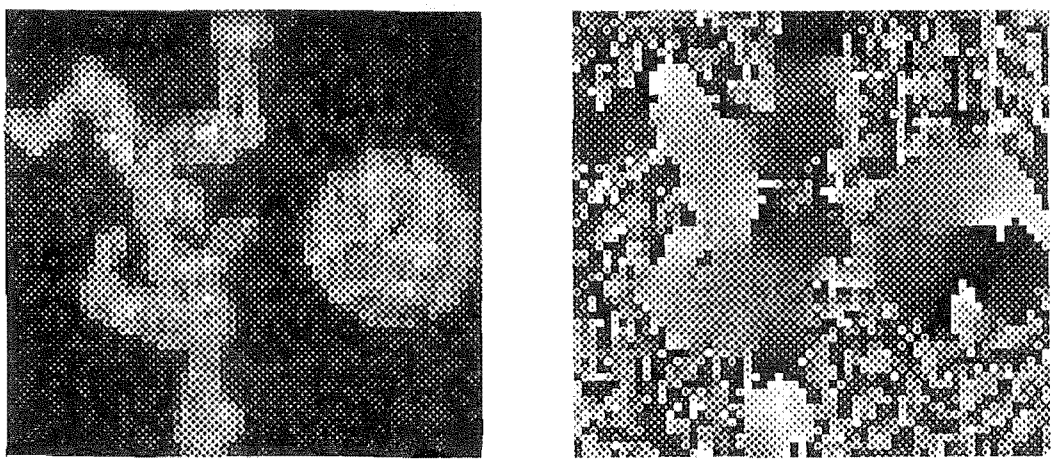


Figure 3.13 Estimates of the image-form of the $f(x, y)$ shown in Figure 3.12(a) generated by the error reduction algorithm: (a) g'_{250}; (b) g'_{700}; (c) g'_{3500}. Amplitudes and phases are displayed as in Figure 3.12.

(a)



(b)



(c)

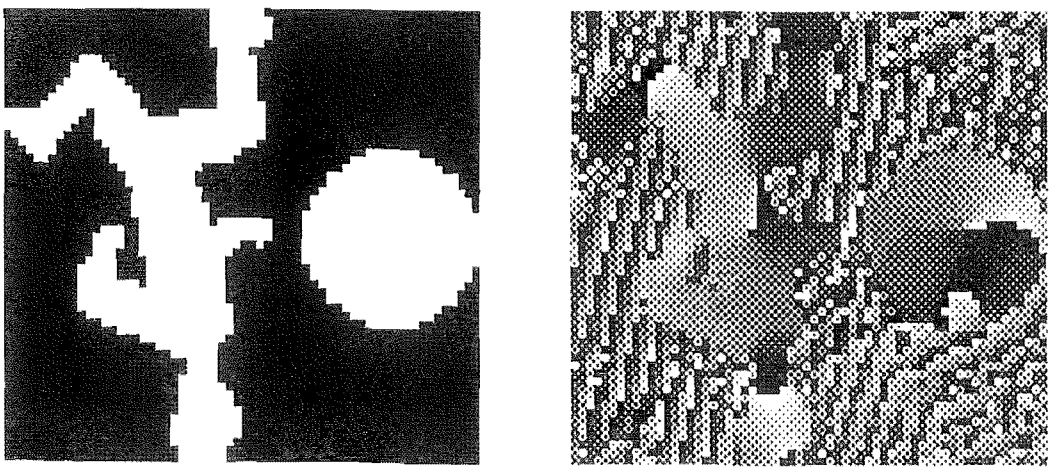


Figure 3.14 Estimates of the image-form of the $f(x, y)$ shown in Figure 3.12(a) generated by the hybrid input-output algorithm: (a) $g'_{250}(x, y)$; (b) $g'_{700}(x, y)$; (c) $g'_{3500}(x, y)$. Amplitudes and phases are displayed as in Figure 3.12.

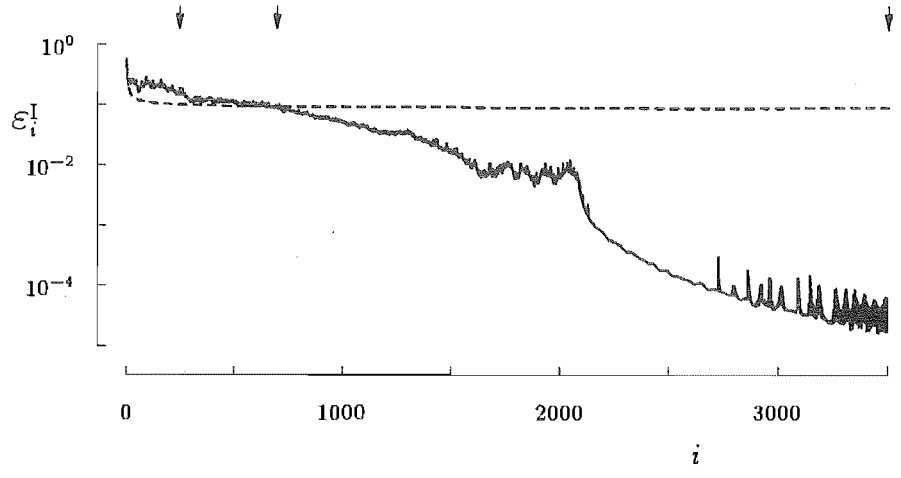


Figure 3.15 Image error $\mathcal{E}_i^I$ plotted against the number of iterations i for both the error reduction example (dashed line) and the hybrid input-output example (solid line). The arrows indicate the iterations at which $g_i'(x,y)$ are displayed in Figures 3.14 and 3.13.

3.5 PHASE RETRIEVAL IN ANTENNA PRACTICE

The algorithm described in the previous section does not represent the only feasible way of solving the Fourier phase problem. In particular, various approaches have been developed for, and adapted to, the radio engineering case. These approaches are discussed in this section.

The *radio engineering phase problem* is: retrieve the copolar aperture field distribution given measured values of only the amplitude of an antenna's copolar radiation pattern. The methods described in Section 3.3 can be employed to perform the required measurements. The measurements are usually, but not necessarily, made in either the Fourier Fresnel region (Sec. 2.1.3.5) or the far field region. The test antenna is assumed directional enough for its radiation pattern to be negligible outside the small angle region (defined in (2.32)). This means that the Fourier transform relationship, expressed by the first equation of (2.57) (or (2.58)), relates the copolar aperture field distribution to the copolar far field (or Fourier Fresnel) pattern.

In keeping with the image processing terminology introduced in Section 3.4, $f(x,y)$ and $F(u,v)$ are defined by

$$f(x,y) = E_{\text{co}}(x,y) \quad \text{and} \quad F(u,v) = \frac{-j\lambda R}{e^{-jkR}} \acute{E}_{\text{co}}(u,v) \tag{3.74}$$

The complex scalar function $f(x,y)$ is called the *copolar aperture field distribution* and, when measurements are made in the far field region, $F(u,v)$ is called the *copolar far field pattern*. The point of the definitions introduced in (3.74) is that, as (2.57) confirms, $f(x,y)$ and $F(u,v)$ are complex scalar distributions related by Fourier transformation.

The support S^f of $f(x,y)$ is, by definition, the region of the x,y plane outside of which $f(x,y)$ vanishes (Sec. 3.4.1.1). The copolar aperture field distribution predicted by the aperture field method (Sec. 2.1.3.3) is zero outside the aperture. Therefore S^f

corresponds to the aperture of the test antenna and is readily available as additional information to aid with the solution of the radio engineering phase problem.

An early, but limited, algorithm for solving the radio engineering phase problem was developed by Davis [1970] and is discussed in Section 3.5.1. Amplitude holography, which is described in Section 3.5.2, is based on the optical holographic principles proposed by Gabor [1948; 1949] and was first investigated in ultrasonic and electromagnetic contexts by Bates and Napier [1971]. Like the Gerchberg-Saxton algorithm (Sec. 3.4.3.1), the Misell algorithm, presented in Section 3.5.3, was originally developed for use in electron microscopy [Misell, 1973a]. A similar algorithm is the plane-to-plane diffraction algorithm which is discussed in Section 3.5.4. All of these methods are compared in Section 3.5.5.

Section 3.5.6 discusses a technique for extrapolating complex measurements of the far field. Although it is not a phase retrieval method, it can be described by a specialization of the basic iterative Fourier transform algorithm. It is described here because aspects of the technique are utilized in the modified Gerchberg-Saxton algorithm (see Sec. 4.7.3.3).

3.5.1 Davis' method

Davis' [1970] method of solving the radio engineering phase problem utilizes a quadratic model for the copolar aperture field phase distribution. The parameters of the model are chosen so that the copolar amplitude pattern predicted by the aperture model best fits the measured copolar amplitude pattern. For simplicity, the method is here described for application to a paraboloidal reflector antenna. Application to other antenna types is straightforward.

The model for the aperture phase deviation $\Delta\psi(x, y)$ (defined in Sec. 3.1) is defined by [Davis, 1970, Sec. IV-D]

$$\Delta\psi(x, y) = \alpha x^2 + \beta xy - \alpha y^2 + \gamma(x^2 + y^2) + ax + by \tag{3.75}$$

This expression includes all terms, less than third order, of a Taylor's series expansion of the actual aperture phase deviation distribution, except that the constant term has been neglected for reasons given in Section 3.2.2. The parameters α and β describe the phase deviation due to astigmatic defects of the reflector (Sec. 3.1). The parameter γ is proportional to the amount by which the feed is axially displaced from the focus of the paraboloid which best fits the reflector's shape. The phase deviation due to this defocusing is approximated by a radially quadratic distribution (Sec. 3.2.2). The parameters a and b represent any linear phase distribution across the aperture.

The estimate of the copolar aperture field distribution $f_e(x, y)$ is defined by

$$f_e(x, y) = f_d(x, y)e^{j\,\Delta\psi(x,y)} \tag{3.76}$$

where $f_d(x, y)$ is the design copolar aperture field distribution and is assumed to be available. When the linear terms are zero (i.e. $a = b = 0$) the associated copolar far field amplitude pattern $|F_e(u, v)|$ can exhibit the following important properties. If $f_d(x, y)$ is point symmetric (see (3.65)) then $|F_e(u, v)|$ is also point symmetric. If $f_d(x, y)$ is circularly symmetric, $|F_e(u, v)|$ has two lines of symmetry which are both described by

$$\tan 2\Theta = \frac{-\beta}{2\alpha} \tag{3.77}$$

where Θ is the angle of the lines from the u axis in an anticlockwise direction. Therefore, the model is invalid when $|F_e(u,v)|$ does not have two lines of symmetry but $f_d(x,y)$ is circularly symmetric.

Davis has also noted that the expected copolar far field amplitude pattern corresponding to a particular set of the parameters $\alpha, \beta, \gamma, a, b$ is exactly the same as that corresponding to $-\alpha, -\beta, -\gamma, a, b$ when $f_d(x,y)$ is conjugate point symmetric (defined in (3.57)). Therefore, from a measured copolar far field amplitude pattern $A_m(u,v)$, it is impossible to tell if the former or the latter set of parameters represents the actual copolar aperture field phase distribution. This accords with it only being possible to retrieve the image-form when solving the Fourier phase problem (Sec. 3.4.2) because both sets of parameters correspond to the same image-form.

The measured amplitude pattern $A_m(u,v)$ may be recorded with the aid of any of the measurement arrangements described in Section 3.3.3.1. The amplitude (or power) pattern of the copolar far field on a regular grid on the u,v plane is required. The field must be oversampled by a factor of at least two (Sec. 3.4.2.1).

The algorithm utilized by Davis minimizes the error E over the parameters α, β, γ, a and b where

$$E = \iint \left[(A_m(u,v))^2 - |F_e(u,v)|^2 \right]^2 du\, dv \tag{3.78}$$

A conjugate gradient iterative method is used to perform the minimization. To resolve the ambiguity in the parameters, it is suggested that the beamwidth of the amplitude pattern be measured after the feed has been moved axially by a known amount. Because the absolute change in γ is known (from the amount by which the feed is moved), the expected beamwidth can be calculated for each of the two possible sets of new parameters. The set of parameters which best predicts the measured beamwidth is then chosen as the correct set.

Davis [1970, Sec. VI-A] has successfully applied this method to a 5.3 m antenna as part of a procedure for correcting the reflector's astigmatism.

3.5.2 Amplitude holography

When a known reference field is added to the far field of a test antenna, each vector component of these fields interfere constructively or destructively depending on their relative phase difference. The amplitude of the sum therefore encodes information about the phase pattern of the test antenna. In what is here called *amplitude holography*, this reasoning is utilized to determine the geometrical defects of the test antenna from measured amplitudes of the combined copolar far field radiated by the test antenna and a reference antenna. The term 'amplitude holography' can be more descriptively replaced by 'amplitude-only microwave Fourier holographic metrology' [cf. Anderson, 1977]. The use of the word 'holography' is discussed in the introduction to this chapter. In contrast to complex holography (Sec. 3.3.3.2) only the amplitude pattern, and not the phase pattern, of the far field is measured directly.

Let $r(x,y)$ be a known reference copolar aperture field distribution existing in the aperture plane of the test antenna. A method of realizing $r(x,y)$ in practice is discussed later in this section. The total copolar aperture field distribution $h(x,y)$ is then given by

$$h(x,y) = f(x,y) + r(x,y) \tag{3.79}$$

The corresponding copolar far field pattern, which is found by Fourier transforming

(3.79), is

$$H(u,v) = F(u,v) + R(u,v) \tag{3.80}$$

The *radiation pattern hologram* [Napier and Bates, 1973] is then defined as the intensity of $H(u,v)$:

$$|H(u,v)|^2 = |F(u,v)|^2 + F^*(u,v)R(u,v) + F(u,v)R^*(u,v) + |R(u,v)|^2 \tag{3.81}$$

The inverse Fourier transform of the radiation pattern hologram is the autocorrelation of the total copolar aperture field distribution:

$$hh(x,y) = ff(x,y) + \tilde{f}(x,y) \odot r(x,y) + f(x,y) \odot \tilde{r}(x,y) + rr(x,y) \tag{3.82}$$

An ideal reference copolar aperture field distribution is a point source

$$r(x,y) = A\,\delta(x - x_0, y) \tag{3.83}$$

where A is its complex amplitude and $(x_0, 0)$ is its position in the x, y plane. For this case (3.82) simplifies to

$$hh(x,y) = ff(x,y) + Af^*(x_0 - x, -y) + A^* f(x + x_0, y) + |A|^2\,\delta(x,y) \tag{3.84}$$

Figure 3.16 compares $h(x,y)$ with $hh(x,y)$. It is apparent from this figure that the $A^* f(x + x_0, y)$ term is spatially isolated from the other terms in (3.84) provided that

$$x_0 > \frac{3D}{2} \tag{3.85}$$

where D is the diameter of the test antenna. When this condition is met, $A^* f(x + x_0, y)$ can be identified by inspection of $hh(x,y)$. Amplitude holography can therefore reconstruct a distribution which is proportional to a shifted version of $f(x,y)$. This is an improvement over being able to compute only the image-form of $f(x,y)$, because the distribution reconstructed by amplitude holography has no ambiguity with regard to its conjugate reflection. Accordingly, any imperfections in the actual aperture distribution have the same locations as those indicated in the reconstructed distribution.

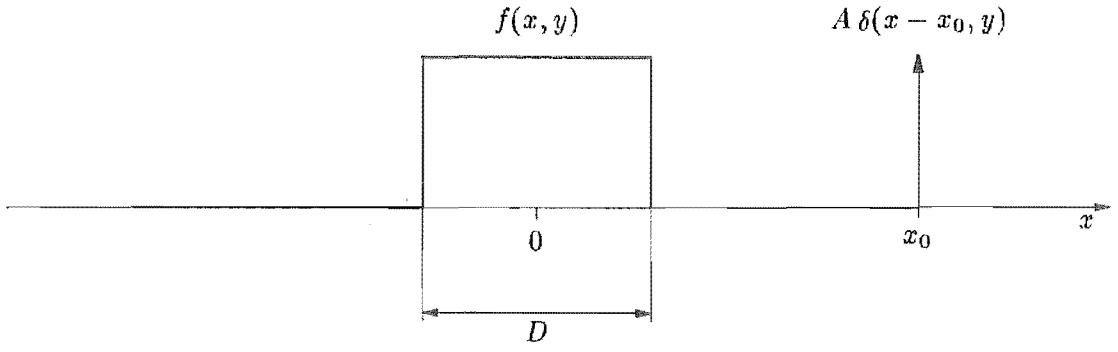
In the following discussion it is assumed that the source, employed for the measurements, is stationary (e.g. located terrestrially or on a geostationary satellite — see Sec. 3.3.3.1). The reference copolar aperture field distribution (3.83) can be realized by considering its copolar far field pattern:

$$R(u,v) = Ae^{j2\pi x_0 u} \tag{3.86}$$

This can be straightforwardly synthesized by keeping the reference antenna pointed towards the source, thereby fixing the value of A, and shifting the phase of the signal from the reference antenna according to $2\pi x_0 u$ [Napier and Bates, 1971]. The variable u depends on the direction in which the test antenna is pointing, relative to the direction of source, and is defined by (2.28). The reference antenna can therefore be of any size and can be positioned anywhere in the vicinity of the test antenna. The radiation pattern hologram is recorded by summing the signal from the test antenna with the phase shifted signal from the reference antenna and measuring the power of the resultant signal.

A consequence of the separation condition (3.85) is that the extent of $h(x,y)$ is twice that of $f(x,y)$ in the x direction. This means that the radiation pattern hologram must be sampled twice as finely, in the u direction, than required for measuring

(a)



(b)

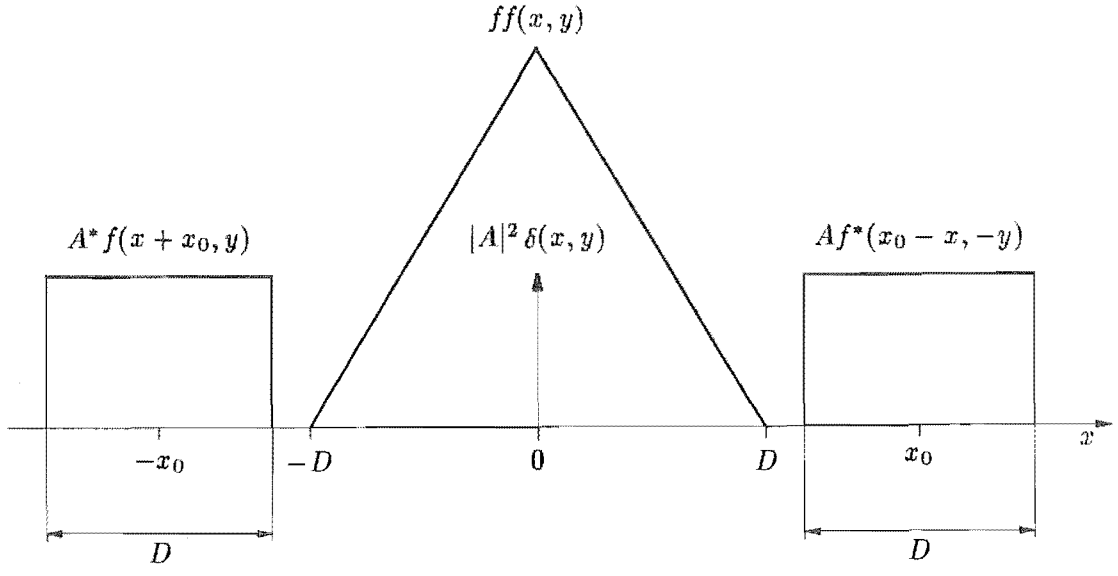


Figure 3.16 Amplitude holography. Cuts along the x axis of (a) $h(x,y) = f(x,y) + A\,\delta(x-x_0,y)$ and (b) $hh(x,y)$ are depicted.

$|F(u,v)|$ alone. Consequently, the measurement time is increased. This situation can be avoided by measuring $|F(u,v)|^2$ separately and then subtracting it from the radiation pattern hologram [Napier and Bates, 1971, Sec. 3.3]. The inverse Fourier transform of the result is given by the right side of (3.84) excluding the first term. The physical implication of this can be seen by removing the curve representing $ff(x,y)$ from the illustration in Figure 3.16(b). This implies that the $A^* f(x+x_0, y)$ term can be shifted closer to the origin while still remaining spatially isolated from the remaining terms, which means that the separation condition (3.85) can be relaxed to $x_0 > D/2$. When $x_0 = D/2$, the extent of $h(x,y)$ equals that of $f(x,y)$, so that their Fourier transform amplitudes can be sampled at the same rate. Therefore, this procedure reduces the measurement time, especially since $|F(u,v)|^2$ can be recorded simultaneously with $|H(u,v)|^2$ by measuring the square of the amplitude of the signal coming directly from the test antenna. A further advantage of measuring $|F(u,v)|^2$, as well as $|H(u,v)|^2$, is that the estimate, generated by amplitude holography, of the copolar aperture field distribution can be verified by comparing its Fourier transform amplitude with $|F(u,v)|$ [Napier and Bates, 1971].

Napier and Bates [1973] have performed laboratory experiments with acoustic antennas. Utilizing amplitude holography, they obtain copolar aperture distributions with an estimated accuracy of better than 5%. Bennett *et al.* [1976] have applied amplitude holography to a 3.66 m diameter microwave paraboloidal reflector. Because their 11 GHz source was located in the Fourier Fresnel region of the test antenna, their estimate of the copolar aperture field had to be adjusted by a quadratic phase term (cf. (2.58)). The adjusted estimate revealed astigmatism in the field radiated by the feed of the antenna. It also indicated that the feed was laterally displaced, which was confirmed by subsequent direct measurement of the feed's position, and that the rms error in the reflector's profile was 0.93 mm, which agreed well with the manufacturer's result of 0.75 mm obtained by template measurements. Bennett *et al.* [1976] also mention that amplitude holography has successfully been applied to a 76 m diameter antenna, utilizing a cosmic radio source.

3.5.3 The Misell algorithm

In radio engineering terminology, the *Misell algorithm* [Misell, 1973a; 1973b; 1973c] requires two copolar amplitude patterns to be measured: one with the antenna in focus and the other with the antenna defocused. The algorithm consists of an iterative error reduction procedure to find a copolar aperture field distribution which is consistent with both copolar amplitude patterns.

The defocusing is achieved by axially displacing either the feed or the subreflector, if one is present. The distance through which the displacement occurs must be noted so that its effect on the copolar aperture field distribution can be calculated (Sec. 3.2.2). The amount of defocus is not critical as long as it reduces the peak gain significantly compared to the noise level on the measured amplitude pattern [Morris, 1985].

The Misell algorithm is a generalization of the basic iterative Fourier transform algorithm (Sec. 3.4.3) and is outlined in Figure 3.17. The operations depicted by the boxes entitled 'transform to ...' account for the effects of defocus on the copolar aperture field distribution, so that $g_i(x,y)$ always represents the focused copolar aperture field distribution. The algorithm is described by the following equations:

$$
\begin{aligned}
\hat{G}_i(u,v) &= \mathrm{FT}\Big\{g_i(x,y)e^{j\,\Delta\psi_{\mathrm{f}}(x,y)}\Big\} \\
\hat{G}'_i(u,v) &= \hat{A}_{\mathrm{m}}(u,v)e^{j\,\mathrm{phase}\{\hat{G}_i(u,v)\}} \\
g'_i(x,y) &= \mathrm{IFT}\Big\{\hat{G}'_i(u,v)\Big\}\,e^{-j\,\Delta\psi_{\mathrm{f}}(x,y)} \\
g_{i+1}(x,y) &= \begin{cases} g'_i(x,y) & \text{for } (x,y)\in S^f \\ 0 & \text{elsewhere} \end{cases}
\end{aligned}
\tag{3.87}
$$

When i is an odd integer, $\hat{A}_{\mathrm{m}}(u,v)$ is the measured copolar amplitude pattern for the focused antenna, so that $\Delta\psi_{\mathrm{f}}(x,y) = 0$. For these iterations, the Misell algorithm is the same as Fienup's error reduction algorithm for complex images (Sec. 3.4.3.3). However, when i is an even integer, $\hat{A}_{\mathrm{m}}(u,v)$ is taken to be the measured copolar amplitude pattern for the defocused antenna, and therefore $\Delta\psi_{\mathrm{f}}(x,y)$ is the aperture phase deviation due to the defocusing. A quadratic phase deviation (3.8) is usually employed [Morris, 1985]. Note, however, that (3.7) is a more exact formulation for $\Delta\psi_{\mathrm{f}}(x,y)$. Convergence of the algorithm is indicated by the normalized rms difference between $|\hat{G}_i(u,v)|$ and $A_{\mathrm{m}}(u,v)$ (i.e. $\mathcal{E}_i^{\mathrm{F}}$ in (3.60)).

The Misell algorithm is usually applied to amplitude patterns which are oversampled by a factor of between one and two [Morris, 1985; 1988; Ellder *et al.*, 1984; Anderson and

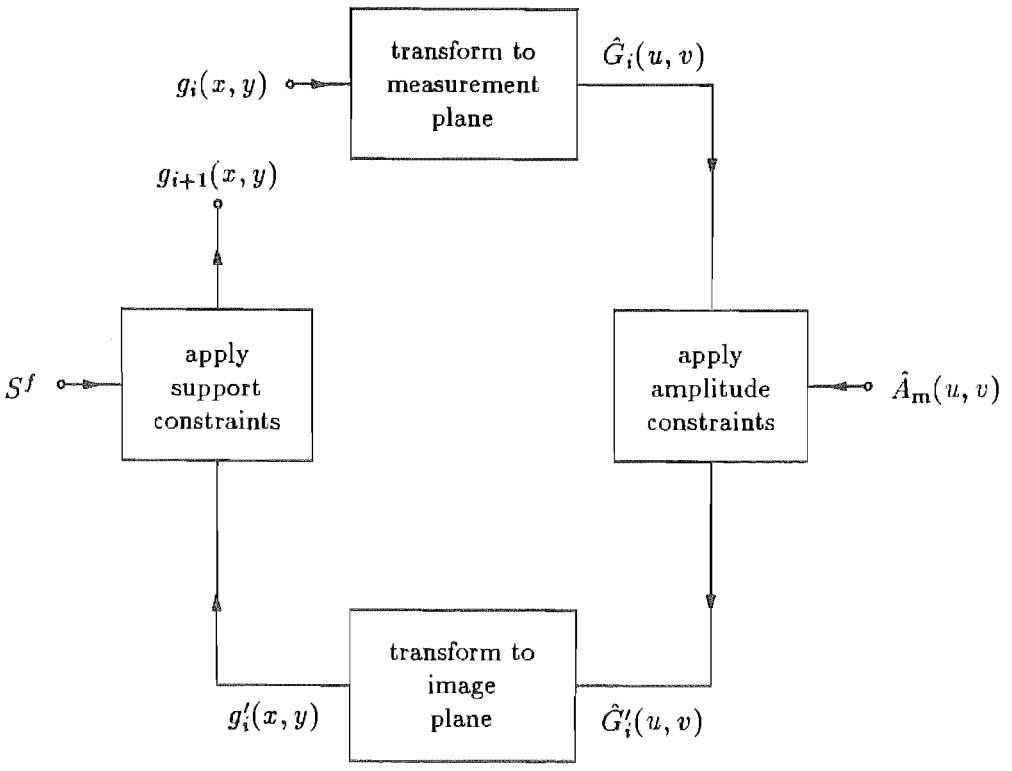


Figure 3.17 A single iteration of the Misell and plane-to-plane diffraction algorithms. When i is an odd integer, $\hat{A}_{\mathrm{m}}(u,v)$ represents one measured amplitude pattern, but when i is an even integer, $\hat{A}_{\mathrm{m}}(u,v)$ represents another measured amplitude pattern.

Sali, 1985]. None of the quoted authors have reported that the algorithm has converged to other than the expected solution. This is in contrast to the Fourier phase problem which requires an amplitude pattern to be oversampled by a factor of at least two, to have a unique solution (Sec. 3.4.2). It is reasonable to expect a smaller sampling factor to be adequate for the Misell algorithm because, whereas the Fourier phase problem is expressed in terms of only a single amplitude pattern, the Misell algorithm requires two amplitude patterns.

The Misell algorithm has been applied to computer simulated antenna patterns by Morris [1985] and Anderson and Sali [1985]. Morris has also derived a relationship between the expected error of the estimated copolar aperture field phase distribution and the noise level in the measured copolar amplitude patterns. From this relationship, and from extensive computer simulations, it is concluded that a peak signal to noise ratio of 50 dB provides enough information for the antenna to be corrected such that its actual gain is within 1% of the desired gain.

Ellder *et al.* [1984] report a test of the Misell algorithm on a 20 m diameter antenna. A cosmic radio source was sensed at 22 GHz and each amplitude pattern was sampled at 32 by 32 points. The focused and defocused amplitude patterns were each measured nine times. Each set of nine measured amplitude patterns was then averaged to reduce the noise level. The geometrical defects predicted by the Misell algorithm were accurate to within an rms error of 0.14 mm. They agreed well with a previous theodolite measurement.

In another antenna measurement performed by Morris *et al.* [1988], an 86 GHz terrestrial source was placed in the Fourier Fresnel region of a 60 m diameter antenna. Each measured amplitude pattern was sampled at 128 by 128 points. The rms repeatability of the computed geometrical defects was around 0.065 mm, being limited mainly by the effects of atmospheric scintillations on the measured amplitude. Because the amplitude pattern measurements were made in the Fresnel region, $\Delta\psi_f(x, y)$ in the first and third equations of (3.87) was appropriately adjusted by a Fresnel quadratic aperture phase term — see (2.58). One-dimensional computer simulations suggest that the sign of the defocusing should be chosen so that the defocus phase term partially cancels the Fresnel phase term [Sali, 1988].

3.5.4 Plane-to-plane diffraction algorithm

The *plane-to-plane diffraction algorithm* was developed by Anderson and Sali [1985] and is based on the Misell algorithm. Both of these algorithms have the same structure, which is depicted in Figure 3.17. The plane-to-plane diffraction algorithm requires measurements of the copolar field amplitude distribution to be made at two different distances from the test antenna. The algorithm attempts to find a copolar aperture field distribution which is compatible with both of the measured field amplitude distributions.

When the measurements are made in the Fourier Fresnel or far field region, the plane-to-plane diffraction algorithm is described by (3.87). However, the first and third equations now represent field transformations between the aperture plane and either of the measurement planes. Comparing with (2.58), $\Delta\psi_f(x, y)$ is seen in (3.87) to be the Fresnel aperture phase term

$$\Delta\psi_f(x, y) = -k\frac{x^2 + y^2}{2R} \tag{3.88}$$

where R is the distance to the sphere over which the measurement is made. Note that, although the angular variation of the copolar field amplitude distribution is measured over a portion of a sphere, the angles are transformed onto the u, v plane by (2.28). As for the Misell algorithm, the copolar amplitude patterns of the Fourier Fresnel or far field are usually oversampled by a factor of between one and two [Anderson and Sali, 1985].

The copolar amplitude measurements can alternatively be carried out in the part of the near field region of the antenna which is outside the Fourier Fresnel region. When measured near to the aperture of the antenna, the copolar field amplitude distribution should be recorded over a plane. The first and third equations of (3.87) must then be replaced by an invertible transformation which relates the aperture field to the field over the measurement plane [Shewell and Wolf, 1968; Anderson and Sali, 1985; Ransom and Mittra, 1971]. The amplitude of the copolar field over the measurement plane must be sampled at closer than $\lambda/2$ in each direction [Anderson and Sali, 1985].

In the original algorithm [Anderson and Sali, 1985], the iterations alternate between the aperture plane and each of the two measurement planes in turn. However, it has been found [Sali and Anderson, 1987a] that the algorithm converges more rapidly if several consecutive iterations involve only one of the measurement planes, with the next several iterations involving only the other measurement plane.

Application of the plane-to-plane diffraction algorithm to computer simulated copolar amplitude distributions indicates that the algorithm can successfully retrieve copolar aperture field distributions [Anderson and Sali, 1985]. The rate of convergence of the algorithm tends to increase with the separation of the measurement planes. However, Sali and Anderson [1987b] have reported an example of the algorithm working successfully when both of the amplitude pattern measurements are made in the far field region. The two measurements are each performed by sampling the copolar far field amplitude pattern at the Nyquist rate, with the samples from the two measurements interleaved. Note that this is not equivalent to a single measurement oversampled by a factor of two (as required by the Fourier phase problem): oversampling occurs in both the u and the v directions, and is therefore equivalent to interleaving four sets of Nyquist spaced samples [cf. Sali and Anderson, 1987b].

Two tests of the plane-to-plane diffraction algorithm have been made by Sali and Anderson [1987a]. In the first, the amplitude pattern of the copolar field was measured at distances 25 m and 70 m from a 3.66 m antenna. In the second test, the amplitude distribution of the copolar field of a 0.45 m antenna was measured over planes 0.31 m and 1.0 m from the aperture plane. For both tests, the plane-to-plane diffraction algorithm produced results which agree well with results obtained from complex holography. It was found that for best convergence the two amplitude measurements should be made to the same resolution and should be accurately registered with respect to each other.

3.5.5 Comparison of methods

In this section the algorithms described in Sections 3.5.1 to 3.5.4 are compared on the basis of how they are applied in practice.

Both amplitude holography and Davis' method require a single measured copolar far field amplitude pattern, oversampled by a factor of at least two. In contrast, the Misell and plane-to-plane diffraction algorithms require two measured amplitude patterns, each oversampled by a factor which need not be as large as two, but must be at least

unity. The latter two algorithms can therefore operate successfully on fewer data than can the former two algorithms.

Since in Davis' algorithm, the copolar aperture field phase distribution is defined by only five parameters, it is only capable of detecting astigmatism and defocusing of the test antenna. Unlike the other algorithms discussed in this section, which have as many parameters as there are samples of the copolar aperture field, Davis' algorithm cannot reveal the effects of panel displacements or other localized defects. Furthermore, it has the disadvantage that it generates an estimate of only the image-form of the copolar aperture field distribution. The associated ambiguity, between the actual copolar aperture distribution and its conjugate reflection, can only be resolved by making a further measurement.

Amplitude holography is a direct method, in the sense that it does not employ an iterative algorithm. It does, however, require a separate reference antenna and phase shifter. Amplitude holography cannot therefore be employed at sites where either of these items is unavailable. In situations where a reference antenna is available, if a phase meter is employed instead of the phase shifter, complex holography (Sec. 3.3.3.2) can be employed. Complex holography has the advantage over amplitude holography of involving fewer samples of the far field and therefore requiring less measurement time. Note that, in concept, a phase shifter can be utilized to construct a crude phase meter: the signal to be measured is added to a phase shifted reference signal, and the shift for which the sum is minimized is recorded.

The Misell algorithm requires the antenna to be defocused between measurements. The defocusing involves physical displacement of either the subreflector or the feed of the test antenna. The subreflector or feed can never be returned to exactly its original position. This has the disadvantage that a feed, say, which was originally in an optimal position, might be displaced by the end of the measurements. The plane-to-plane diffraction algorithm, on the other hand, does not require any physical tampering with the antenna to perform the measurements. It does, however, require measurements to be made at two different distances from the antenna, at least one of which is in the near field region. The measurements therefore require either two separate sources or a single movable source. This may well be inconvenient. Furthermore, for both the Misell and the plane-to-plane diffraction algorithms, the two measured amplitude patterns must be recorded carefully enough to be accurately registered with respect to each other [McCormack *et al.*, 1989].

3.5.6 Far field extrapolation

In complex holography (Sec. 3.3.3.2), the measurements of the complex copolar far field pattern are made over a limited range of angles. Ideally, for the appropriate use of the DFT operator (Sec. 3.4.1.4), the field outside this range of angles should be negligible. In practice, this is often untrue, however, so that the measured copolar far field pattern is significantly truncated. This limits the resolution obtainable for the copolar aperture field distribution.

It is worth noting in passing that the entire copolar far field pattern $F(u, v)$ can, in principle, be determined from accurate information about only a finite portion of $F(u, v)$. The reason for this is that, because the copolar aperture field distribution $f(x, y)$ is of finite extent, $F(u, v)$ is analytic [Papoulis, 1975]. Therefore, the whole of $F(u, v)$ can be expressed as a Taylor series expansion about any given point. However, this does not constitute a practical approach because the determination of the Taylor

series coefficients is extremely sensitive to noise. In order to successfully extrapolate $F(u,v)$, it is necessary to find a method which incorporates the image constraints into the extrapolation procedure.

The problem, then, is to extrapolate the data $F_{\mathrm{m}}(u,v)$, obtained from a measurement of $F(u,v)$ over a finite portion of the u,v plane, when $f(x,y)$ is known to be of finite extent. This is similar to the Fourier phase retrieval problem because in both problems the whole of $F(u,v)$ is required to be retrieved from partial information about it. In fact, the algorithm developed by Gerchberg [1974] and Papoulis [1975] to solve the extrapolation problem is a special form of the basic iterative Fourier transform algorithm (Sec. 3.4.3). The i^{th} iteration can be described by

$$\begin{aligned} G_i(u,v) &= \mathrm{FT}\{g_i(x,y)\} \\ G_i'(u,v) &= \begin{cases} F_{\mathrm{m}}(u,v) & \text{for } (u,v) \in S^{F_{\mathrm{m}}} \\ G_i(u,v) & \text{elsewhere} \end{cases} \\ g_i'(x,y) &= \mathrm{IFT}\{G_i'(u,v)\} \\ g_{i+1}(x,y) &= \begin{cases} g_i'(x,y) & \text{for } (x,y) \in S^f \\ 0 & \text{elsewhere} \end{cases} \end{aligned} \tag{3.89}$$

where $S^{F_{\mathrm{m}}}$ is the region of the u,v plane over which the measurements are made.

Gerchberg [1974] proves that the image error, defined in (3.71), must decrease as the number of iterations increases. Papoulis [1975] analyzes the effect of noise and aliasing on the algorithm's performance. Convergence of the algorithm is faster the more accurate an estimate of the aperture support S^f is available. However it is better to overestimate S^f because the algorithm fails when too small a support is employed [Gerchberg, 1974].

The algorithm has been applied to the case of extrapolating a measured copolar far field pattern for complex holography by Rahmat-Samii [1984]. Both computer simulated measured data and data from measurements of a 64 m diameter antenna have been successfully extrapolated.

3.6 SUMMARY

An antenna is designed to produce a radiation pattern which meets its specifications. If it does not meet these specifications, and provided the characteristics of the environment and the feed are as expected, there must be geometrical defects of the antenna itself. There are many forms of geometrical defect: random variations in reflector shape associated with a finite manufacturing tolerance, misaligned panels, displaced feed or subreflector, and astigmatic deflections of the main reflector due to gravitational loading.

A geometrical defect which is an appreciable fraction of a wavelength produces an appreciable deviation in the copolar aperture field phase. The same defect tends to have an insignificant effect on the amplitude of the copolar aperture field distribution. From knowledge of the aperture phase deviations, an estimate of the geometrical defects can be calculated. For an antenna designed to have a uniform copolar phase distribution across the aperture, the aperture phase deviation is equal to the actual copolar aperture field phase distribution. An approximately radially quadratic copolar aperture field phase distribution suggests an axially displaced feed or subreflector. A panel shaped

area of phase advance in the copolar aperture field distribution suggests the panel immediately behind the area is misaligned. The geometrical defects associated with any given copolar aperture field phase distribution can be inferred by tracing rays backwards from the aperture plane to the feed.

There are many methods of determining the geometrical defects of an antenna. One way is to measure the geometry directly and compare it with the design geometry. Other ways rely on first obtaining an estimate of the phase distribution of the copolar aperture field and then inferring the geometrical defects from this. The copolar aperture field phase distribution can be measured directly or obtained by measuring either the copolar Fresnel or far field pattern. When the complex copolar far field pattern is measured, the copolar aperture field distribution can be computed by inverse Fourier transformation. When measurements of only the amplitude pattern of the copolar far field are made, phase retrieval algorithms can be invoked to retrieve the copolar aperture field distribution. This thesis is concerned with the latter approach.

The problem of retrieving the copolar aperture field distribution $f(x, y)$, when only the copolar far field amplitude pattern $|F(u, v)|$ is available, is called the Fourier phase problem. Because $f(x, y)$ is of finite extent, z-transform theory can be employed to show that in general the phase problem has a unique solution, provided $|F(u, v)|$ is oversampled by a factor of at least two. A solution is here understood to be any copolar aperture field distribution having the same image-form as $f(x, y)$. A practical algorithm for solving the Fourier phase problem is the basic iterative Fourier transform algorithm which is described in Section 3.4.3. An important feature of this algorithm is that it can incorporate any additional information about $f(x, y)$. The Gerchberg-Saxton algorithm and Fienup's algorithms are examples of the basic iterative Fourier transform algorithm.

The radio engineering phase problem is broader than the Fourier phase problem. It is the problem of retrieving the copolar aperture field distribution from measured spatial distributions of the copolar amplitude of the field radiated by an antenna. Several methods are available for solving the radio engineering phase problem. Davis' method is only suitable for copolar aperture fields which have quadratic phase distributions. Amplitude holography requires a separate reference antenna. In the Misell and plane-to-plane diffraction algorithms, two separate measurements are required, with either the subreflector, feed, or the measuring source being moved between the measurements. The modified Gerchberg-Saxton algorithm, introduced in the next chapter, does not require a reference antenna and operates on only a single measured copolar amplitude pattern of either the Fresnel or far field.

CHAPTER 4

THE MODIFIED GERCHBERG-SAXTON ALGORITHM: EVALUATION BY COMPUTER SIMULATION.

This chapter introduces the modified Gerchberg-Saxton algorithm which is an adaptation of the original Gerchberg-Saxton algorithm (Sec. 3.4.3.1) suited to help solve the radio engineering phase problem (Sec. 3.5). Ideally, the inputs to the original Gerchberg-Saxton algorithm should be the actual copolar far field amplitude pattern and the actual copolar aperture field amplitude distribution, which are related by Fourier transformation. The algorithm outputs an estimate of the actual copolar aperture field distribution. The obvious way to implement the Gerchberg-Saxton algorithm, in the radio engineering situation, is to provide it with a measured copolar far field amplitude pattern and a measured copolar aperture field amplitude distribution. However, as pointed out by Morris [1985] and Anderson and Sali [1985], it is inconvenient to make both of these measurements because they involve different techniques: measuring the copolar far field amplitude pattern requires a standard amplitude pattern measurement (Sec. 3.3.3.1), while measurement of the copolar aperture field amplitude distribution requires a planar near field scanning technique (Sec. 3.3.2). Furthermore, Anderson and Sali [1985] performed computer simulations comparing the plane-to-plane diffraction algorithm, the Misell algorithm and the Gerchberg-Saxton algorithm, showing that the Gerchberg-Saxton algorithm stagnated within a few iterations, while the other algorithms continued converging and produced more accurate results.

The approach taken when the modified Gerchberg-Saxton algorithm was developed is somewhat different. Rather than making measurements in the aperture plane, the design copolar aperture field amplitude distribution, along with the measured copolar far field amplitude pattern, are used as inputs to the algorithm. It is appropriate to utilize the design copolar aperture field amplitude distribution because, as reasoned in Section 3.2.1, the difference in amplitude between the design and actual copolar aperture field distributions is expected to be relatively small when it is due to geometrical defects. However, due to causes other than geometrical defects, the design copolar aperture field amplitude distribution may not be reliably close to the actual copolar aperture field amplitude distribution. These causes include amplitude deviations in the feed pattern and unavoidable errors in the analysis technique employed for determining the design copolar aperture field distribution from the design reflector geometry. Therefore, the modified Gerchberg-Saxton algorithm has been made to rely heavily on the measured copolar far field amplitude pattern, while ensuring that the amplitude of the estimated copolar aperture field distribution is not forced to equal the design copolar aperture field amplitude distribution. This is achieved by making the final iterations of the modified Gerchberg-Saxton algorithm equivalent to Fienup's error reduction algorithm for complex images (Sec. 3.4.3.3). This approach can be viewed as solving the Fourier phase problem as posed in (3.41), using the design

copolar aperture field amplitude distribution only as an aid for convergence. As intimated in Section 3.4.2, for the Fourier phase problem to possess a unique solution, it is crucial to oversample, by a factor of at least two, the copolar far field amplitude pattern. Furthermore, oversampling by a factor of at least two helps the Gerchberg-Saxton algorithm to converge [Gardenier *et al.*, 1986b; 1986c]. Other ways of improving the convergence of the algorithm are to invoke Gerchberg's variant of the Gerchberg-Saxton algorithm (Sec. 3.4.3.2) [Milner *et al.*, 1987] or to incorporate aspects of Fienup's hybrid input-output algorithm (Sec. 3.4.3.3) [Bates *et al.*, 1987]. In both of these forms of the modified Gerchberg-Saxton algorithm, the design data are employed for the initial iterations but are discarded for the final iterations.

A similar approach has been explored by the group at the University of Sheffield. Sali and Anderson [1987b] describe Fienup's error reduction algorithm for complex images and find that it does not produce satisfactory results when applied to computer simulated data. The discussion in Section 3.4.3.3 suggests that this is to be expected. Anderson *et al.* [1988] do, however, get the error reduction algorithm to converge satisfactorily by utilizing an inaccurately measured copolar phase pattern in the first iteration of the algorithm. This algorithm can therefore be used to improve an estimate of the phase pattern. Both Fienup's hybrid input-output and error reduction algorithms are utilized by McCormack and Anderson [1988]. They start their algorithm with the design copolar aperture field distribution, which is discarded after the first iteration. They applied the algorithm to experimentally obtained measured data and found that the resulting estimate of the copolar aperture field phase distribution was noisier than, but very similar to, the copolar aperture field phase distribution generated by complex holography. As with the algorithm mentioned in the next sentence, the design copolar aperture field amplitude distribution was in the form of a rough estimate of the copolar aperture field amplitude distribution. McCormack *et al.* [1989] have developed an algorithm which is a combination of Fienup's algorithms and the Gerchberg-Saxton algorithm. This algorithm and the Misell algorithm were applied to experimentally measured data. The results from the Misell algorithm were inferior because the two measured amplitude patterns were not aligned properly. The combined Fienup and Gerchberg-Saxton algorithms produced results similar to those from complex holography. They were not sensitive to the choice of design copolar aperture field amplitude distribution. Anderson *et al.* [1989] have modified this algorithm so that it can be used with data obtained by measuring the field on a plane in the near field region. The Fourier transform operator is replaced by an invertible transformation which relates the aperture field to the field over the measurement plane (cf. the plane-to-plane diffraction algorithm in Sec. 3.5.4).

A suggested procedure for applying the modified Gerchberg-Saxton algorithm in practice is discussed in Section 4.1. A generalized computer model which simulates relevant parts of the practical procedure is described in Section 4.2. Section 4.3 defines error measures which indicate how well the modified Gerchberg-Saxton algorithm performs when it is applied to data from a particular computer model. The algorithm has a variety of forms differing according to their constraints, which are discussed in Section 4.4. It is in this section that the form of the modified Gerchberg-Saxton algorithm with which this thesis is mainly concerned is defined. A worked example is presented in Section 4.5. Aspects of measuring the far copolar field amplitude pattern are discussed in Section 4.7. Section 4.8 presents the results of applying the algorithm to many sets of computer simulated data, while Section 4.9 indicates uses for the algorithm other than phase retrieval.

4.1 PRACTICAL IMPLICATIONS OF THE ALGORITHM

In this chapter the performance of the modified Gerchberg-Saxton algorithm is evaluated by applying it to antennas, fields and measurements which are all computer simulated. Before taking the computer simulation approach, however, it is instructive to consider how the algorithm might be applied to real-world antennas, fields and measurements. Section 4.1.1 suggests how the modified Gerchberg-Saxton algorithm could be employed for detecting and correcting geometrical defects of an antenna. Possible ways of dealing with the ambiguity inherent in the modified Gerchberg-Saxton algorithm in the real world of antenna engineering are discussed in Section 4.1.2.

4.1.1 Procedure

Section 3.1 presents a general procedure for determining whether or not an antenna has geometrical defects, for correcting them and for confirming that the corrections are successful. It does not specify what method is to be employed to determine the geometrical defects. This section discusses in more detail how the modified Gerchberg-Saxton algorithm might play a key role in revealing the geometrical defects.

To determine if the antenna geometry requires adjustment, the copolar far field amplitude pattern is measured and compared to its specifications. If the specifications are given along a single cut, the measurements need only be made along that cut. If the measured pattern does not meet its specifications and if other factors, such as the environment or the feed characteristics, are considered not to be at fault, the modified Gerchberg-Saxton algorithm can be invoked to determine the geometrical defects of the antenna.

The inputs to the modified Gerchberg-Saxton algorithm are a measured copolar far field amplitude pattern and the design copolar aperture field amplitude distribution. Because the algorithm is implemented on a computer, these inputs are required to be in the form of a sampled Fourier transform amplitude and a sampled image respectively (Sec. 3.4.1.2). In order to straightforwardly utilize the DFT operator, the number and spacing of the sample points in the u, v plane and the x, y plane must be chosen to satisfy the conditions set down in (3.37). As for complex holography (Sec. 3.3.3.2), it is assumed that the far field pattern of the antenna is negligible outside the small angle region.

For reasons given in Section 4.7.2, the *measured copolar far field amplitude pattern*, denoted by $A_{\mathrm{m}}(u, v)$, must be oversampled by a factor of at least two. Therefore, from (3.22), the spacing of the sample points in both the u and v directions should be less than $1/(2D)$, where D is the largest dimension of the antenna's aperture. Methods for measuring the two-dimensional copolar far field amplitude pattern are discussed in Section 3.3.3.1.

The *design copolar aperture field distribution* $f_{\mathrm{d}}(x, y)$, which is defined in Section 3.1, can be obtained in a number of ways, depending on the information available about the design of the antenna. Some antennas are designed to produce a particular desired aperture field distribution [e.g. Galindo-Israel *et al.*, 1987]. For these antennas $f_{\mathrm{d}}(x, y)$ can be the copolar component of this desired aperture field distribution. Another way of determining $f_{\mathrm{d}}(x, y)$ is to utilize the feed characteristics and reflector geometry specified by the design. An analysis technique, such as one of those described in Section 2.1, is employed to compute the aperture field distribution expected to be produced by the specified feed characteristics and the design reflector geometry. The copolar aperture

field distribution derived from such an analysis is then taken to be $f_{\mathrm{d}}(x,y)$. All is not lost even if no design information is available because $f_{\mathrm{d}}(x,y)$ can be a guess at the ideal copolar aperture field distribution. As intimated in Sections 2.2.5 and 2.1.3.3, this ideal distribution typically incorporates the effects of any subreflector or feed blockage, possesses an amplitude taper and is negligible outside the aperture. The size and shape of the aperture can usually be adequately estimated by inspection of the antenna.

The modified Gerchberg-Saxton algorithm assumes that the *actual copolar aperture field distribution* $f_{\mathrm{a}}(x,y)$ is related to the *actual copolar far field pattern* $F_{\mathrm{a}}(u,v)$ by Fourier transformation (Sec. 3.3.3.2). From the energy conservation theorem for Fourier transforms [Bates and McDonnell, 1989, p. 24], the energies, as defined in (3.18), of $f_{\mathrm{a}}(x,y)$ and $F_{\mathrm{a}}(u,v)$ are equal. Since $|f_{\mathrm{d}}(x,y)|$ and $A_{\mathrm{m}}(u,v)$ are approximations to $|f_{\mathrm{a}}(x,y)|$ and $|F_{\mathrm{a}}(u,v)|$ respectively, $A_{\mathrm{m}}(u,v)$ must be normalized so that it has the same energy as $|f_{\mathrm{d}}(x,y)|$. This normalization implies that only the relative amplitude pattern of the copolar far field need be measured. Furthermore the distance R, which appears in (2.57), between the test antenna and the source need not be known.

The output from the modified Gerchberg-Saxton algorithm is the *estimated copolar aperture field distribution* $f_{\mathrm{e}}(x,y)$. As outlined in Section 3.2, the phase of this distribution can be utilized to determine the geometrical defects of the antenna. These geometrical defects can then be corrected by the means suggested in Section 3.1. The copolar far field amplitude pattern of the corrected antenna can then be measured to check that it does conform to its specifications.

Throughout this thesis, the copolar amplitude pattern measurement is conveniently assumed to be made in the far field region. However, the measured copolar amplitude pattern of the Fourier Fresnel field can alternatively be used as the input $A_{\mathrm{m}}(u,v)$ to the modified Gerchberg-Saxton algorithm. By comparing (2.57) with (2.58), it is then apparent that the estimated copolar aperture field distribution $f_{\mathrm{e}}(x,y)$ generated by the algorithm, must be multiplied by $e^{jk(x^2+y^2)/(2R)}$ before it can become an estimate of the actual copolar aperture field distribution.

4.1.2 Ambiguities

Section 3.4.2.4 concludes that, in general, the two-dimensional Fourier phase problem, as posed in (3.41), has one, and only one, solution. In the derivation of this result, it is implicitly assumed that the Fourier transform amplitude is known accurately, i.e. $A_{\mathrm{m}}(u,v) = |F_{\mathrm{a}}(u,v)|$. If the estimate $f_{\mathrm{e}}(x,y)$ generated by the modified Gerchberg-Saxton algorithm is such that $|F_{\mathrm{e}}(u,v)| = A_{\mathrm{m}}(u,v)$ then $f_{\mathrm{e}}(x,y)$ is a solution to the Fourier phase problem. Because this solution is necessarily unique, $f_{\mathrm{e}}(x,y)$ must have the same image-form (defined in (3.38)) as $f_{\mathrm{a}}(x,y)$.

Because the position of the aperture in the x,y plane is known, the position of the support of $f_{\mathrm{e}}(x,y)$ is also known. Therefore $f_{\mathrm{e}}(x,y)$ cannot be a translated form of $f_{\mathrm{a}}(x,y)$. However, $f_{\mathrm{e}}(x,y)$ can still be related to $f_{\mathrm{a}}(x,y)$ in the following ways (cf. (3.38)):

$$f_{\mathrm{e}}(x,y) = f_{\mathrm{a}}(x,y)e^{j\psi_0} \quad \text{or} \quad f_{\mathrm{e}}(x,y) = \tilde{f}_{\mathrm{a}}(x,y)e^{j\psi_0} = f_{\mathrm{a}}^*(-x,-y)e^{j\psi_0} \tag{4.1}$$

where ψ_0 is an arbitrary real number. Recall, from Section 3.4.2, that the tilde implies that $\tilde{f}_{\mathrm{a}}(x,y)$ is the conjugate reflection of $f_{\mathrm{a}}(x,y)$. The uniform phase term $e^{j\psi_0}$ is unimportant in radio engineering contexts and can be ignored for reasons given in Section 2.3.1.2. In practice, the modified Gerchberg-Saxton algorithm converges to either $f_{\mathrm{a}}(x,y)e^{j\psi_0}$ or $\tilde{f}_{\mathrm{a}}(x,y)e^{j\psi_0}$ plus an error term which is dependent upon how accurately

$A_m(u,v)$ approximates $|F_a(u,v)|$. It is assumed, in the following discussion, that the modified Gerchberg-Saxton always converges well enough for its output $f_e(x,y)$ to approximate either $f_a(x,y)e^{j\psi_0}$ or $\tilde{f}_a(x,y)e^{j\psi_0}$. Furthermore, the effect of the uniform phase term is ignored by assuming that $\psi_0 = 0$. Note that the previous two sentences imply that either $f_e(x,y)$ or $\tilde{f}_e(x,y)$ approximate $f_a(x,y)$.

If $f_a(x,y)$ is conjugate point symmetric (defined in (3.57)), $f_c(x,y) \approx f_a(x,y)$ and there is no ambiguity. However, in general, $f_a(x,y)$ is not conjugate point symmetric. This presents a serious problem in practice, because it is not obvious whether $f_e(x,y)$ or $\tilde{f}_e(x,y)$ is the estimate of $f_a(x,y)$. This is of practical importance because the required geometrical corrections are different for each case. Various approaches to resolving this ambiguity problem are now presented.

Any known asymmetries in the copolar aperture field distribution $f_a(x,y)$ can help decide whether $f_e(x,y)$ or $\tilde{f}_e(x,y)$ is the estimate of $f_a(x,y)$. For example, if $|f_d(x,y)|$ is not point symmetric (defined in (3.65)), the estimate of $f_a(x,y)$ is taken to be whichever of $f_e(x,y)$ or $\tilde{f}_e(x,y)$ has the closest amplitude distribution to $|f_d(x,y)|$. If the position and/or extent of some of the geometrical misalignments are known, their effect on the copolar aperture field phase distribution can be computed. If this effect is manifest in one but not the other of $f_e(x,y)$ or $\tilde{f}_e(x,y)$, then that distribution is taken to be the estimate of $f_a(x,y)$. Asymmetries can also be purposely introduced into $f_a(x,y)$. This can be done, for example, by placing a piece of microwave absorbing material on part of the reflector surface, or by purposely displacing some part of the antenna structure, before making the measurement of the copolar far field amplitude pattern. A disadvantage with adding absorbing material is that it obscures any shape defects of the part of the reflector over which it is placed. It is of course essential that any purposely displaced part of the antenna be accurately replaced afterwards.

To avoid making alterations to the antenna before its amplitude pattern is measured, the ambiguity can be resolved in the following way. Initially assume that $f_e(x,y)$ is the estimate of $f_a(x,y)$, and make appropriate corrections to the geometry of the antenna. If subsequent measurement of the radiation pattern reveals that it is now within specifications, then no more need be done. However, if the radiation pattern is worse than it was before the changes were made, it must be concluded that $\tilde{f}_e(x,y)$ is the estimate of $f_a(x,y)$. The initial geometrical changes must thus be undone. New changes, this time based on $\tilde{f}_e(x,y)$, must then be made.

Making changes to the antenna geometry can be a protracted business and therefore be expensive, especially if a large number of individual panels need be adjusted. This may well make the method described in the previous paragraph unattractive because there is a 50% chance that the geometry of the antenna has to be changed for a second time.

Another way of resolving the ambiguity between $f_e(x,y)$ and $\tilde{f}_e(x,y)$ is based on making initial changes to the antenna geometry, measuring the ensuing pattern and then undoing the initial changes. In this approach, the initial changes should be as few as possible and limited to those which can be easily reversed. Examples of readily reversible changes are moving the feed along its axis, or placing a metal sheet on the reflector surface to produce the effect of displacing a panel. Alternatively, microwave absorbing material can be placed over a part of the reflecting surface. After making these initial changes, the copolar aperture field distribution $f_c(x,y)$ becomes

$$f_c(x,y) = f_a(x,y)\kappa(x,y) \tag{4.2}$$

where $\kappa(x,y)$ represents the effects of the changes on the original copolar aperture

field distribution. An estimate of $\kappa(x,y)$ can be calculated from knowledge of the changes that have been made. For example, if the axial position of the feed in a paraboloidal antenna is changed, $\kappa(x,y) = e^{j\,\Delta\psi(x,y)}$ where $\Delta\psi(x,y)$ is the aperture phase deviation defined in (3.7). When utilizing microwave absorbing material, $|\kappa(x,y)|$ is small over the projection, onto the x,y plane, of the region of the main reflector covered by the material. Note that there are several easily implementable forms for $\kappa(x,y)$. Since either $f_e(x,y)$ or $\tilde{f}_e(x,y)$ must be the estimate of $f_a(x,y)$ then $f_c(x,y)$ must be approximated by either $f_{c1}(x,y)$ or $f_{c2}(x,y)$, where

$$\begin{aligned} f_{c1}(x,y) &= f_e(x,y)\kappa(x,y) \\ \text{or}\quad f_{c2}(x,y) &= \tilde{f}_e(x,y)\kappa(x,y) \end{aligned} \tag{4.3}$$

It is worth emphasizing that if $f_{c1}(x,y)$ and $f_{c2}(x,y)$ do not differ significantly from each other, there is no ambiguity problem. Therefore, if they do differ significantly, $|F_{c1}(u,v)|$ must differ significantly from $|F_{c2}(u,v)|$. Because both $F_{c1}(u,v)$ and $F_{c2}(u,v)$ can be computed, they can be compared to find the points in the u,v plane at which their amplitudes differ most. Measurements of $|F_c(u,v)|$ at these points can then be made to resolve the ambiguity: if $|F_c(u,v)|$ is closest to $|F_{c1}(u,v)|$, $f_e(x,y)$ must be the estimate of $f_a(x,y)$; if $|F_c(u,v)|$ is closest to $|F_{c2}(u,v)|$, $\tilde{f}_e(x,y)$ must be the estimate of $f_a(x,y)$. Because the measurements of $|F_c(u,v)|$ need only be made at a small number of points in the u,v plane, they can be made relatively rapidly. Note that $F_{c1}(u,v)$ and $F_{c2}(u,v)$ can be calculated before any geometrical changes are made. Therefore, the form of $\kappa(x,y)$ can be chosen to provide a large and easily detectable difference between $|F_{c1}(u,v)|$ and $|F_{c2}(u,v)|$. Once the ambiguity is resolved and the initial changes are undone, the final changes, based on the appropriate estimate of $f_a(x,y)$, can be implemented.

As pointed out by McCormack and Anderson [1988], the ambiguity can often be straightforwardly resolved when the copolar amplitude pattern is measured in the Fourier Fresnel region. This is because, as intimated at the end of Section 4.1.1, when measurements are made in the Fourier Fresnel region, $f_a(x,y)$ is the actual copolar aperture field distribution multiplied by a radially quadratic phase term $e^{-jk(x^2+y^2)/(2R)}$. If phase$\{f_a(x,y)\}$ is dominated by this known quadratic phase term, $f_a(x,y)$ can immediately be distinguished from $\tilde{f}_a(x,y)$. It may then be possible to determine which of $f_a(x,y)$ or $\tilde{f}_a(x,y)$ is most nearly approximated by $f_e(x,y)$.

4.2 COMPUTER MODELLING OF REFLECTOR ANTENNAS

In this chapter, the modified Gerchberg-Saxton algorithm is applied to computer generated data. The algorithm is then evaluated by comparing the results it produces with additional computer generated data. Both of these sets of data are obtained with the aid of a generalized computer model of the design process, the aperture field, the far field and the process utilized to measure the copolar far field amplitude pattern. This computer model is described in this section.

The computer model approach to evaluation of algorithms which operate on measured far field patterns is not new. For example, Rahmat-Samii [1984] and Morris [1985] have employed computer modelling techniques to evaluate complex holography and the Misell algorithm respectively. In the model employed by Morris [1985], the DFT operator (Sec. 3.4.1.4) is invoked to compute a sampled copolar far field pattern generated by a specified sampled copolar aperture field distribution. Rahmat-Samii [1984] invokes a polar coordinate form of the Fourier transform operator when simulating the

far field pattern. While this allows an analytic expression for the copolar far field pattern to be derived, it is limited to copolar aperture field distributions which are circularly symmetric. In another model employed by Rahmat-Samii [1985], the shape of the antenna's reflectors and the characteristics of the feed are specified, so that physical optics/Jacobi-Bessel and geometrical theory of diffraction formulations can be utilized to simulate the far field radiation pattern. The model utilized in this chapter invokes the DFT operator to relate samples of the copolar aperture field distribution to samples of the copolar far field distribution. This technique was chosen because of its computational simplicity.

The different steps involved in developing the generalized model are discussed in the following sections. Section 4.2.1 describes two different design copolar aperture field distributions and their far fields patterns. The model permits the simulation of copolar aperture field deviations produced by those types of geometrical defect of the antenna which are of most practical significance. These simulated deviations and the actual copolar aperture field distributions which result from them are discussed in Section 4.2.2. Modelling the inaccuracies inherent in measurement of the copolar far field amplitude pattern is the concern of Section 4.2.3. Finally, Section 4.2.4 describes an extension to the model which simulates depolarization effects.

It is important to realize that, while attempts have been made to make the computer model reasonably realistic, it is not crucial for the model to be realistic. The main purpose of the model is to test the operation of modified Gerchberg-Saxton algorithm with many sets of data, corresponding to a wide variety of geometrical defects and measurement inaccuracies. Although the ultimate test of the algorithm is to apply it to measured far field amplitude patterns of real antennas, the simulations associated with the computer model provide a valuable indication of the applicability and limitations of the algorithm.

4.2.1 Design fields

Each of the computer simulations presented in this chapter uses one of two design copolar aperture field distributions which are referred to as 'design 1' and 'design 2' respectively. Before describing these designs, however, a few comments are made which are relevant to the computer model as a whole.

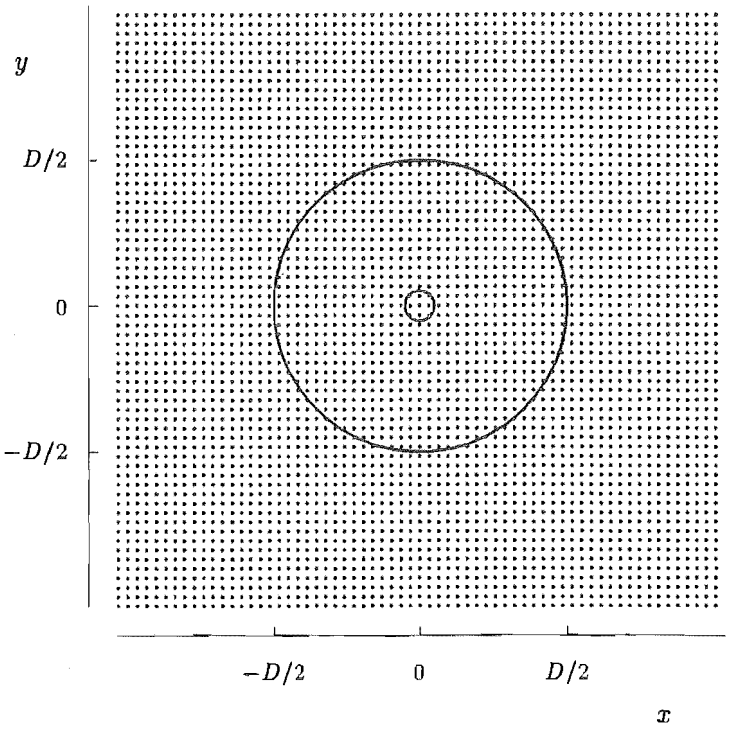
Throughout this chapter all copolar aperture field distributions and copolar far field patterns are represented by arrays of 64 by 64 samples, the positions of which are indicated in Figure 4.1. In the aperture plane, the sample spacings in the x and y directions, Δ_x and Δ_y respectively, are equal. An antenna with a circular aperture of diameter D is modelled. Except in Section 4.7.2, the sample spacings in the aperture plane are such that

$$D = 31\Delta_x = 31\Delta_y \tag{4.4}$$

From (3.22) and the first equation of (3.37), the sampling factors α_u and α_v in the u, v plane are each 64/31. This implies that the far field patterns are oversampled by a factor of greater than 2.

Any point in the aperture plane is identified by its Cartesian coordinates (x, y), where $(0, 0)$ is coincident with the centre of the aperture. However, because many of the aperture field distributions are circularly symmetric, it is also convenient to be able to refer to a point by its *normalized polar coordinates* $(\rho; \phi)$ which are related to (x, y)

(a)



(b)

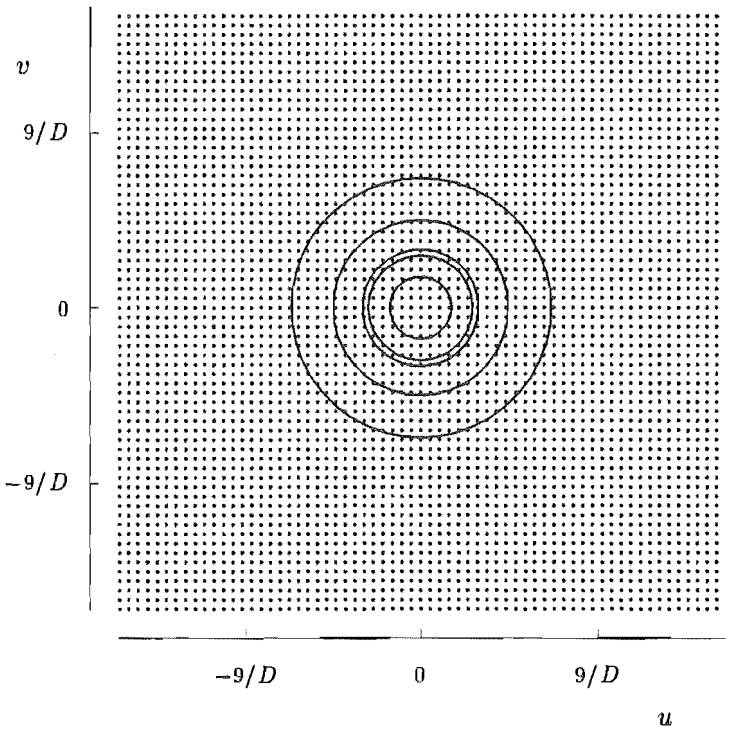


Figure 4.1 Sample points in (a) the x, y plane and (b) the u, v plane. The sample points are represented by dots. The circles in (a) indicate the perimeter of the physical aperture and the subreflector. The circles in (b) are the loci of the first five nulls of the radiation pattern resulting from design 1. The dashed lines indicate the cuts along which $f(x, y)$ and $F(u, v)$ are graphed in many figures in this chapter.

by

$$\begin{aligned} x &= \rho \frac{D}{2} \cos\phi \\ y &= \rho \frac{D}{2} \sin\phi \end{aligned} \tag{4.5}$$

Note that the ρ is distance from the centre of the aperture normalized with respect to the radius of the aperture.

Both of the design copolar aperture field distributions described in this section model the copolar aperture field distributions of Cassegrain antennas (Sec. 2.3.2), which are typical high gain reflector antennas. Both design copolar aperture field distributions are circularly symmetric and incorporate the effects of edge taper as well as subreflector blockage (Sec. 2.2.5). The subreflector diameter is taken to be one tenth that of the aperture [cf. Rudge *et al.*, 1982, p. 165]. The effects of the struts on the aperture field are ignored in the designs, but are modelled in the actual copolar aperture field distributions (Sec. 4.2.2). In both designs the phase distribution of the design copolar aperture field is uniform and taken to have a value of zero radians.

Design 1 is based on a circularly symmetric Gaussian aperture distribution [Rudge *et al.*, 1982, p. 44; Morris, 1985; Lamb and Olver, 1986]. It simulates the copolar aperture field distribution produced by a balanced feed (Sec. 2.3.1.2), possessing a Gaussian radiation pattern, illuminating a prototype Cassegrain antenna (Sec. 2.3.2). It is worth noting that hybrid-mode feeds having radiation patterns of appropriate Gaussian form can be readily fabricated [James, 1980]. The design copolar aperture field distribution $f_{\mathrm{d}}(x, y)$, for design 1, is defined by

$$f_{\mathrm{d}}(x,y) = \begin{cases} e^{-1.725\rho^2} & \text{for } 0.1 \leq \rho \leq 1 \\ 0 & \text{elsewhere} \end{cases} \tag{4.6}$$

which is similar to the distribution utilized in the computer model of Morris [1985].

As mentioned in Section 2.3.2, almost any circularly symmetric copolar aperture field distribution can be obtained from a balanced feed illuminating a suitably shaped subreflector and main reflector. The reflectors can be shaped to produce a copolar aperture field distribution which is tapered at both the edge and the centre of the aperture. The edge taper reduces spillover (Sec. 2.2.5), while the centre taper reduces the blockage effect of the subreflector [Dijk and Maanders, 1968]. A suitable copolar aperture field distribution, proposed by James [1980], is defined by

$$f_{\mathrm{d}}(x,y) = \begin{cases} 1 - 0.82e^{-4(1-\rho)} - 0.82e^{-8\rho} & \text{for } 0.1 \leq \rho \leq 1 \\ 0 & \text{elsewhere} \end{cases} \tag{4.7}$$

This design distribution comprises design 2, which therefore corresponds to a shaped Cassegrain antenna.

Design 1 and design 2 are depicted in Figures 4.2 and 4.3 respectively. Each figure shows the samples of $f_{\mathrm{d}}(x, y)$ which lie along the diagonal cut indicated in Figure 4.1(a). Position along this diagonal is identified by the parameter ξ which is defined by

$$\xi = \sqrt{2}\, x = -\sqrt{2}\, y \tag{4.8}$$

The reason for taking a cut along this diagonal is that it intercepts the majority of the aperture field deviations discussed in Section 4.2.2. The design copolar far field

pattern $F_d(u,v)$ is the Fourier transform of $f_d(x,y)$ and is computed by discrete Fourier transforming the samples representing $f_d(x,y)$. In both Figures 4.2 and 4.3 the samples representing $|F_d(u,v)|$ are graphed along a portion of the u axis, which is indicated in Figure 4.1(b). The values of $|F_d(u,v)|$ are expressed in decibels relative to $|F_d(0,0)|$. Although $|F_d(u,v)|$ does not represent a power level, its square is proportional to the power of the copolar component of the far field, as intimated by (2.43). Therefore, any copolar amplitude pattern, say $|F_d(u,v)|$, can be expressed in decibels as a fraction of a constant amplitude value, say $|F_d(0,0)|$, as $20\log_{10}(|F_d(u,v)|/|F_d(0,0)|)$. Usually an amplitude pattern is normalized to its peak value. However, when two amplitude patterns are being compared, both can be normalized to the peak of just one of the patterns. In graphs, such as these shown in Figures 4.2 and 4.3, the points representing the amplitudes, or phases, of the samples are connected by line segments to visually relate the sample values to each other.

In the remainder of this section, two effects of sampling, namely aliasing and the picket fence effect, are discussed with the aid of Figure 4.4. Consider the set of samples representing $f_d(x,y)$ for design 1. This set of samples is here referred to as the sampled $f_d(x,y)$ and its discrete Fourier transform is called the sampled $F_d(u,v)$. The cuts through the sampled $f_d(x,y)$ and the sampled $F_d(u,v)$ are represented by dots in Figure 4.4. Note that they are also depicted in Figure 4.2. In contrast, what is here called the continuous $f_d(x,y)$ is defined by (4.6) assuming that x and y are allowed to vary continuously. Its continuous Fourier transform is here called the continuous $F_d(u,v)$. Cuts through the continuous $f_d(x,y)$ and the continuous $F_d(u,v)$ are represented by solid curves in Figure 4.4.

As expected, the sampled $f_d(x,y)$ is exactly equal to the continuous $f_d(x,y)$ at the points (x,y) corresponding to the sample points in the aperture plane. However, it is apparent from Figure 4.4(b) that the sampled $F_d(u,v)$ accurately represents the continuous $F_d(u,v)$ near the centre of the u,v plane, but that these samples are less accurate towards the edge. This inaccuracy is due to aliasing, which is discussed in Section 3.4.1.3. It reveals a weakness in the model, because, when using the discrete Fourier transform operator, representations of the copolar far field pattern, such as $F_d(u,v)$, are only accurate near their centre. For this reason, only the centre portions of amplitude patterns are usually plotted in the figures throughout this chapter. Note that complex holography (Sec. 3.3.3.2) also suffers from the effects of aliasing because it too utilizes the DFT operator as an approximation to the Fourier transform operator. Because satisfactory results have been obtained from complex holography, aliasing is not expected to be a serious weakness of the model utilized in this chapter.

A further observation from Figure 4.4 is that the sample points do not necessarily lie at the extrema of the continuous distributions. Therefore plots such as those shown in Figures 4.2 and 4.3 do not always faithfully indicate the maxima and minima of the corresponding continuous distributions. This is called the *picket fence effect* [Stanley *et al.*, 1984, Sec. 10-1] because it is as if the continuous distribution is seen through the gaps between the palings of a picket fence.

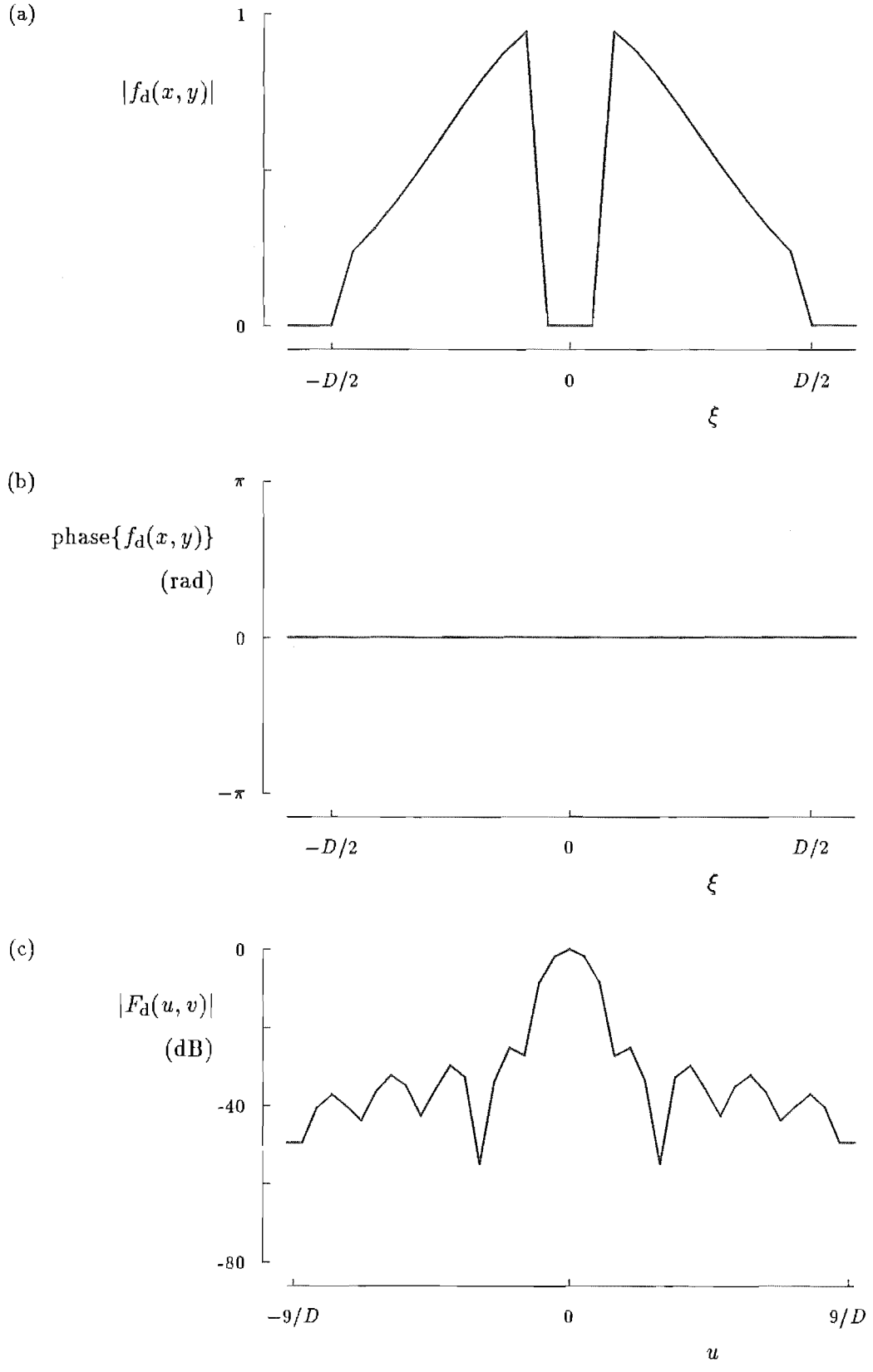


Figure 4.2 Design 1: (a) design copolar aperture field amplitude distribution; (b) design copolar aperture field phase distribution; (c) design copolar far field amplitude pattern. The variable ξ, which identifies position along a diagonal in the aperture plane, is defined in (4.8).

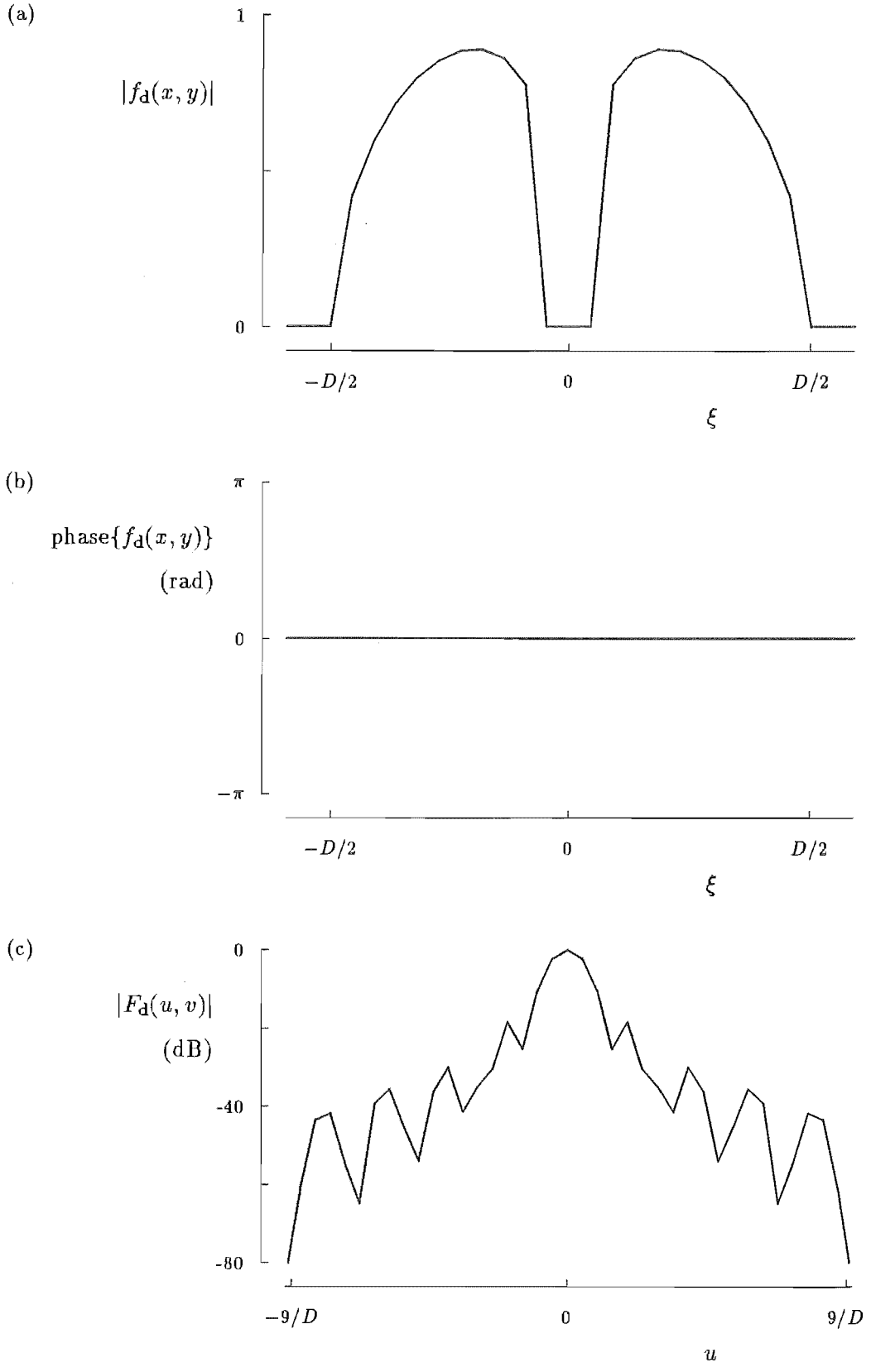


Figure 4.3 Design 2: (a) design copolar aperture field amplitude distribution; (b) design copolar aperture field phase distribution; (c) design copolar far field amplitude pattern. The variable ξ, which identifies position along a diagonal in the aperture plane, is defined in (4.8).

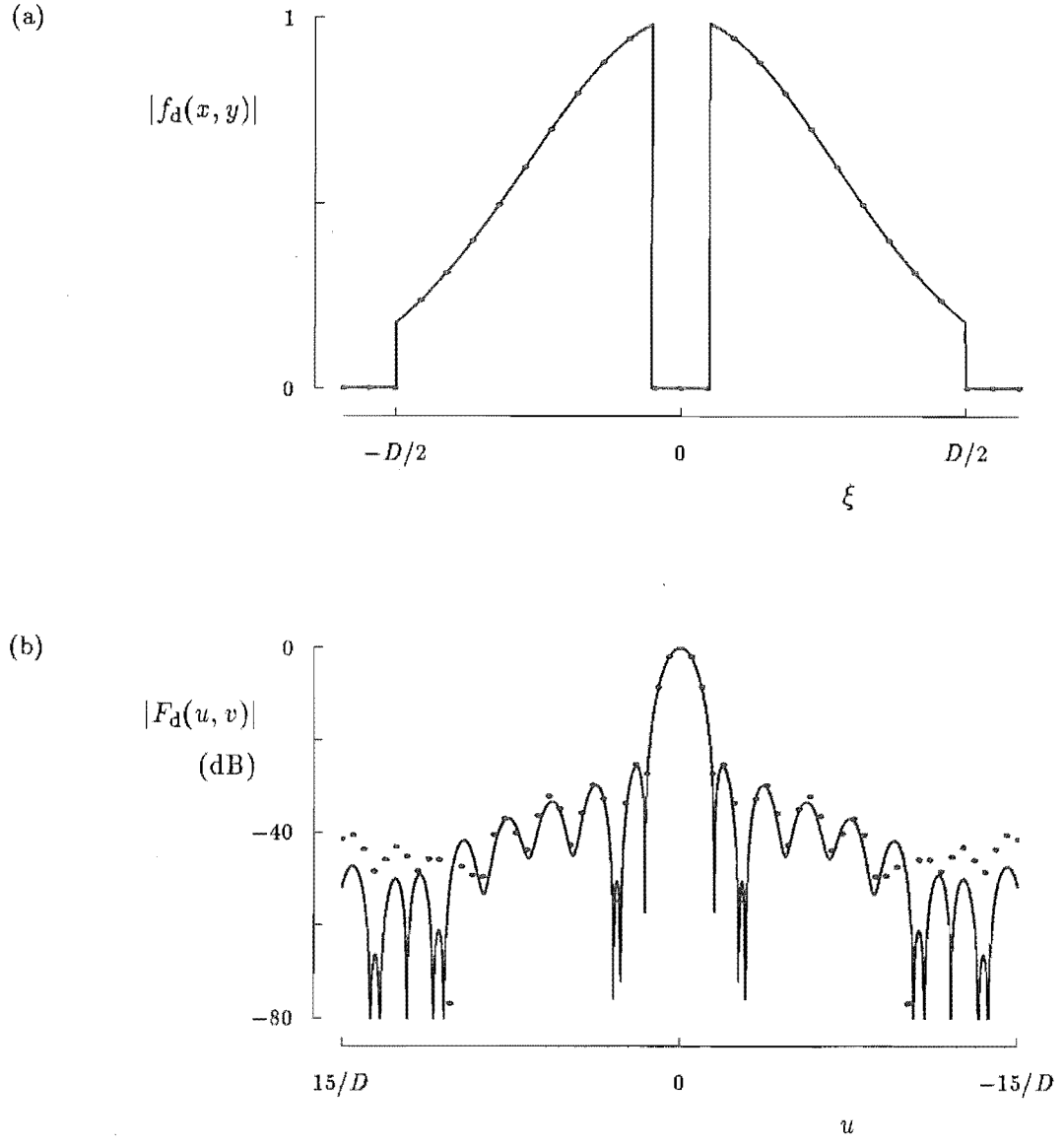


Figure 4.4 Comparison of the continuous and sampled fields for design 1: (a) design copolar aperture field distribution; (b) design copolar far field amplitude pattern. The solid curves represent the continuous design fields. The dots in (a) represent the sampled copolar aperture field distribution, while the dots in (b) represent the discrete Fourier transform of the sampled aperture field distribution. The variable ξ, which identifies position along a diagonal in the aperture plane, is defined in (4.8).

4.2.2 Field deviations

As defined in Section 3.1, field deviations are any differences between the actual and the design fields. In this section, field deviations due to many different causes are described and modelled. The actual copolar aperture field distribution is modelled by

$$f_{\mathrm{a}}(x,y) = [|f_{\mathrm{d}}(x,y)| + \Delta a(x,y)]e^{j\,\Delta\psi(x,y)} + n(x,y) \tag{4.9}$$

where $f_{\mathrm{d}}(x,y)$ is the design copolar aperture field distribution defined in Section 4.2.1. The real valued distribution $\Delta\psi(x,y)$ represents aperture phase deviations which can be straightforwardly related to geometrical defects of the antenna. The real valued distribution $\Delta a(x,y)$ represents easily modelled aperture amplitude deviations such as that caused by a feed whose radiations pattern is narrower than specified by the design. The effect of scattering from struts and other extraneous antenna features is modelled by the complex valued distribution $n(x,y)$. $\Delta\psi(x,y)$, $\Delta a(x,y)$ and $n(x,y)$ are further discussed in the following paragraphs.

The aperture phase deviations caused by geometrical defects of the antenna are analysed in Section 3.2. As intimated in Section 3.1, two common kinds of geometrical defect are a displaced feed and a displaced panel of the main reflector. Referring to (3.3) and (3.8), the aperture phase deviation $\Delta\psi(x,y)$ is here approximated by [cf. Rahmat-Samii, 1985; Morris, 1985]

$$\Delta\psi(x,y) = \psi_{\mathrm{quad}}\rho^2 + \psi_{\mathrm{pan}}\,P(x,y,\rho_{\min},\rho_{\max},\phi_{\min},\phi_{\max}) \tag{4.10}$$

where the real numbers ψ_{quad} and ψ_{pan} characterize the amount of defocus and panel displacement respectively. The region of the aperture plane affected by the displaced panel is identified by $P(x,y,\rho_{\min},\rho_{\max},\phi_{\min},\phi_{\max})$ where

$$P(x,y,\rho_{\min},\rho_{\max},\phi_{\min},\phi_{\max}) = \begin{cases} 1 & \text{for } \rho_{\min} \le \rho \le \rho_{\max},\ \phi_{\min} \le \phi \le \phi_{\max} \\ 0 & \text{elsewhere} \end{cases} \tag{4.11}$$

The normalized area Ω_{pan} of the displaced panel is

$$\Omega_{\mathrm{pan}} = \frac{\iint P(x,y,\rho_{\min},\rho_{\max},\phi_{\min},\phi_{\max})\,dx\,dy}{\iint_{S^{\mathrm{aper}}} dx\,dy} \tag{4.12}$$

where S^{aper} is the *aperture support*. Note that S^{aper} includes the region of the aperture plane which is affected by subreflector blockage. This is in contrast to the support $S^{f_{\mathrm{d}}}$ of $f_{\mathrm{d}}(x,y)$ which, as intimated by (4.6) and (4.7), does not include the region of the aperture plane corresponding to the subreflector. In the majority of computer simulations presented in this chapter, a single displaced panel is modelled. However, the size of the panel differs for different simulations. Table 4.1 and Figure 4.5 indicate the different positions and sizes of panel modelled in the computer simulations presented in this chapter. Note from Table 4.1 that the position of each panel can be uniquely identified by its value of Ω_{pan}. Rahmat-Samii [1985] points out that there is an alternative to interpreting $P(x,y,\rho_{\min},\rho_{\max},\phi_{\min},\phi_{\max})$ as identifying a single displaced panel. Consider a group of adjacent panels located with respect to each other such that the region of the aperture plane affected by their displacement can be identified by $P(x,y,\rho_{\min},\rho_{\max},\phi_{\min},\phi_{\max})$. Assuming that each panel in the group is displaced by an equal amount, the aperture phase deviation due to the displaced panels can be

$\rho_{\min}$	$\rho_{\max}$	$\phi_{\min}$ (deg)	$\phi_{\max}$ (deg)	Ω_{pan}
0.5	0.5	120	120	0.0
0.5	0.629	120	130	0.004
0.5	0.629	120	140	0.008
0.5	0.758	120	140	0.019
0.5	0.951	120	140	0.036
0.5	1.000	120	150	0.063
0.5	1.000	120	180	0.120
0.5	1.000	120	240	0.251

Table 4.1 Details of the different panel positions and sizes modelled in this chapter. In each row, the values of $\rho_{\min}$, $\rho_{\max}$, $\phi_{\min}$ and $\phi_{\max}$ identify the position of a panel, as defined by (4.11), while the value of Ω_{pan} is the normalized area of the panel and is defined by (4.12).

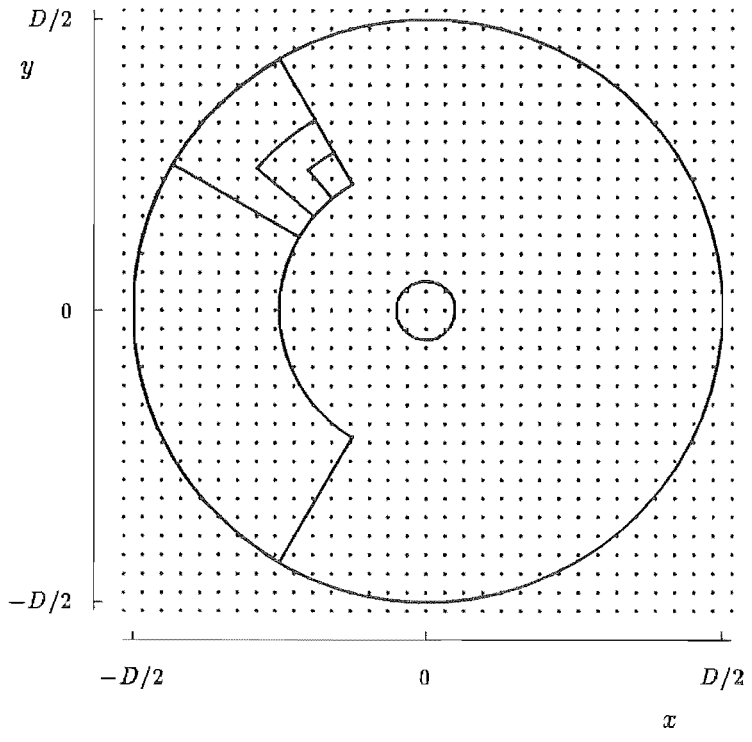


Figure 4.5 Positions of panels modelled in computer simulations. The dots represent the sample points in the aperture plane and the circles indicate the perimeters of the aperture and subreflector. The remaining closed curves indicate the perimeter of the panels identified in Table 4.1 by $\Omega_{\text{pan}} = 0.04$, 0.019, 0.063 and 0.251 respectively.

described by the second term on the right side of (4.10). Different values of Ω_{pan} then correspond to groups consisting of different numbers of panels.

Not all aperture field deviations are caused by geometrical defects. For example, the radiation pattern of the feed may be narrower than is specified by the design. Therefore $f_{\text{a}}(x,y)$ is more tapered than $f_{\text{d}}(x,y)$. This increased tapering is here modelled by

$$\Delta a(x,y) = \begin{cases} -2\,\tau_{\text{quad}}\,\rho^2 + \tau_{\text{quad}} & \text{for } 0.1 \le \rho \le 1 \\ 0 & \text{elsewhere} \end{cases} \tag{4.13}$$

where τ_{quad} characterizes the deviations, associated with increased tapering, between $|f_{\text{d}}(x,y)|$ and $|f_{\text{a}}(x,y)|$. Note that the subreflector blockage is not affected by the increased tapering. Equation (4.13) can also model the situation in which $|f_{\text{d}}(x,y)|$ is guessed (cf. Sec. 4.1.1).

In the definition of $f_{\text{d}}(x,y)$ (Sec. 4.2.1), the effect of subreflector blockage is modelled by setting $f_{\text{d}}(x,y)$ to zero in the region of the aperture plane affected by the blockage. Blockage due to the struts is not included in the model because they often have widths of the order of a wavelength and therefore do not cast a well defined shadow [Rudge *et al.*, 1982, p. 146]. However, other effects of struts are modelled here. Consider a transmitting antenna. Any radiation impinging on a strut is scattered. For a cylindrical strut, this scattering is concentrated in particular directions, which depend on the position and orientation of the strut [Rusch *et al.*, 1982]. However, by employing struts of suitably irregular shape, the scattering can be dispersed over a wide range of directions [Matsunaka *et al.*, 1981]. The effect of such struts can be modelled by replacing the struts with an equivalent source distribution in the aperture plane [Ko *et al.*, 1984]. Accordingly, the effect of irregularly shaped struts is here modelled by setting

$$n(x,y) = \begin{cases} \tau_{\text{ran}}\,[\text{ran}(x,y) + j\,\text{ran}(x,y)] & \text{for } (x,y) \in S^{\text{aper}} \\ 0 & \text{elsewhere} \end{cases} \tag{4.14}$$

where $\text{ran}(\cdot,\cdot)$ is a uniformly distributed pseudo random distribution, independent from sample to sample, with a standard deviation of 1.0. Each time it is invoked, $\text{ran}(\cdot,\cdot)$ is a different distribution. The real number τ_{ran} is proportional to the amplitude of the radiation scattered from the struts.

Note that (4.9) through (4.13) model $f_{\text{a}}(x,y)$, which is characterized by the values of ψ_{quad}, ψ_{pan}, Ω_{pan}, τ_{ran} and τ_{quad}. Each of these parameters has a default value of zero and therefore assumes this value, for any particular model, unless another value is specified. Figures 4.6 to 4.9 illustrate the effect of these parameters on both $f_{\text{a}}(x,y)$ and the actual copolar far field amplitude pattern $|F_{\text{a}}(u,v)|$ relative to $|F_{\text{d}}(0,0)|$. Some of the values of ψ_{quad}, ψ_{pan}, Ω_{pan}, τ_{ran} and τ_{quad} are chosen to be unrealistically large for some of the distributions of $f_{\text{a}}(x,y)$ graphed in these figures so that the effects of the aperture field deviations can be clearly recognized. It can be seen from Figures 4.6 to 4.9 that, as intimated in Section 2.2.5, the general effect of the aperture phase deviations on $|F_{\text{a}}(u,v)|$ is to lower its peak value, widen its main beam and raise the levels of its sidelobes. Complex noise in $f_{\text{a}}(x,y)$ also raises the levels of the sidelobes of $|F_{\text{a}}(u,v)|$. The effect of increased tapering in $|f_{\text{a}}(x,y)|$ is to widen the main beam, but lower the sidelobe levels, of $|F_{\text{a}}(u,v)|$.

Although not actually implemented here, the model of $f_{\text{a}}(x,y)$ can be straightforwardly extended to cover other causes of copolar aperture field deviation, such as lateral feed displacement and astigmatic and random shape defects of the reflector surfaces.

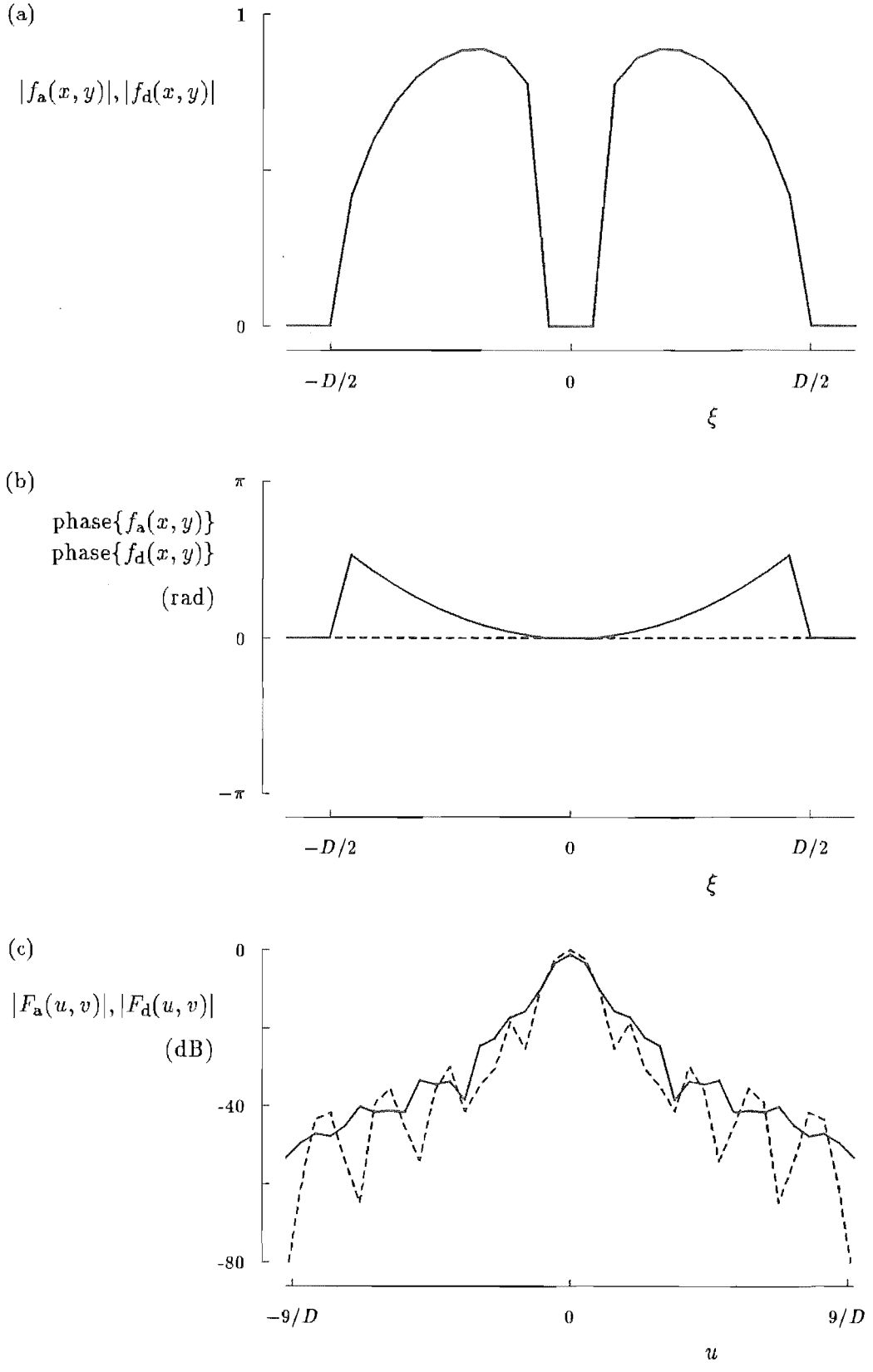


Figure 4.6 Actual copolar fields when $\psi_{\text{quad}} = 2.0$ rad.: (a) copolar aperture field amplitude distribution; (b) copolar aperture field phase distribution; (c) copolar far field amplitude distribution. The solid curves represent the actual copolar fields while the dashed curves represent the design copolar fields associated with design 2. The variable ξ, which identifies position along a diagonal in the aperture plane, is defined in (4.8).

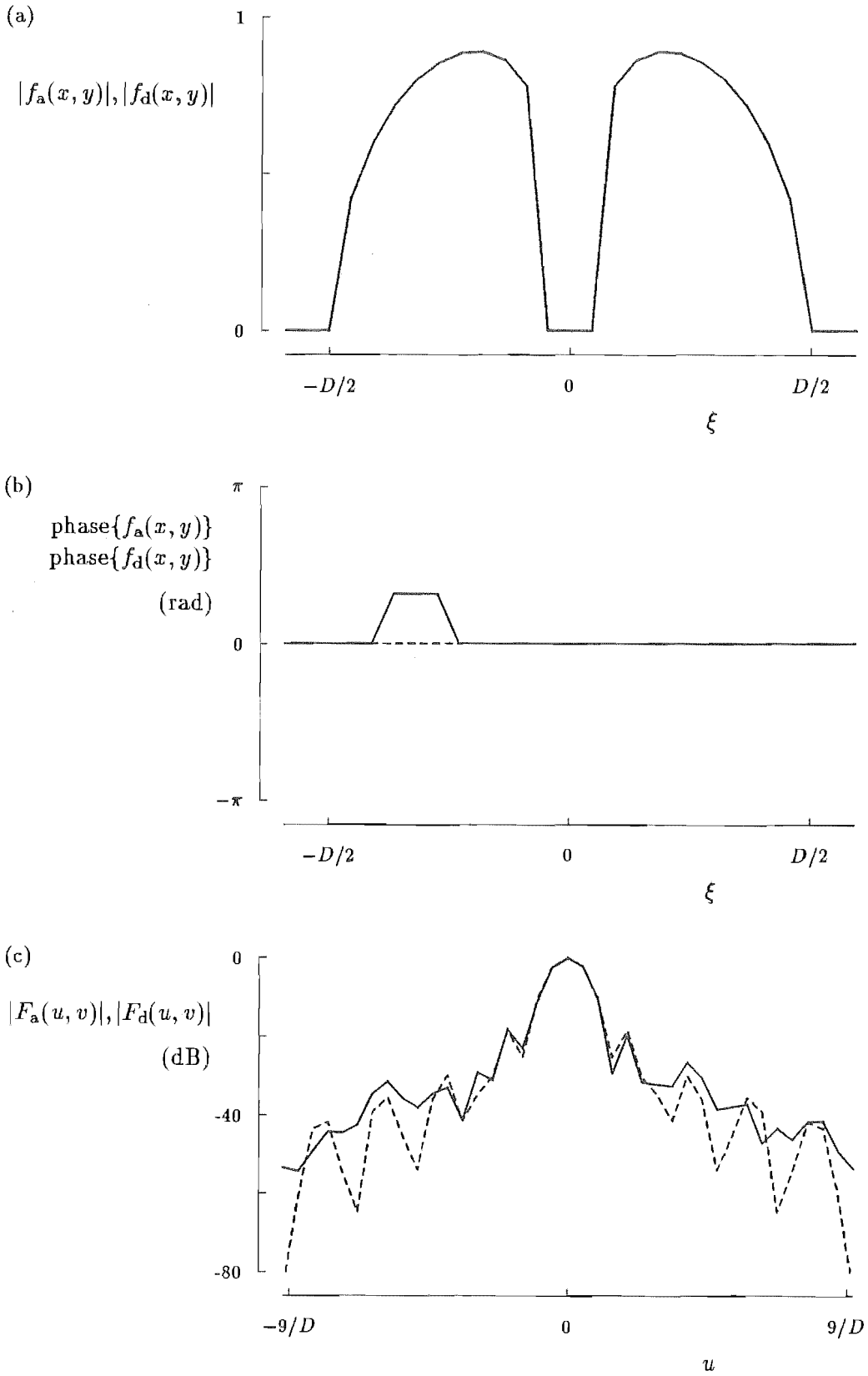


Figure 4.7 Actual copolar fields when $\psi_{pan} = 1.0$ rad. and $\Omega_{pan} = 0.02$: (a) copolar aperture field amplitude distribution; (b) copolar aperture field phase distribution; (c) copolar far field amplitude distribution. The solid curves represent the actual copolar fields while the dashed curves represent the design copolar fields associated with design 2. The variable ξ, which identifies position along a diagonal in the aperture plane, is defined in (4.8).

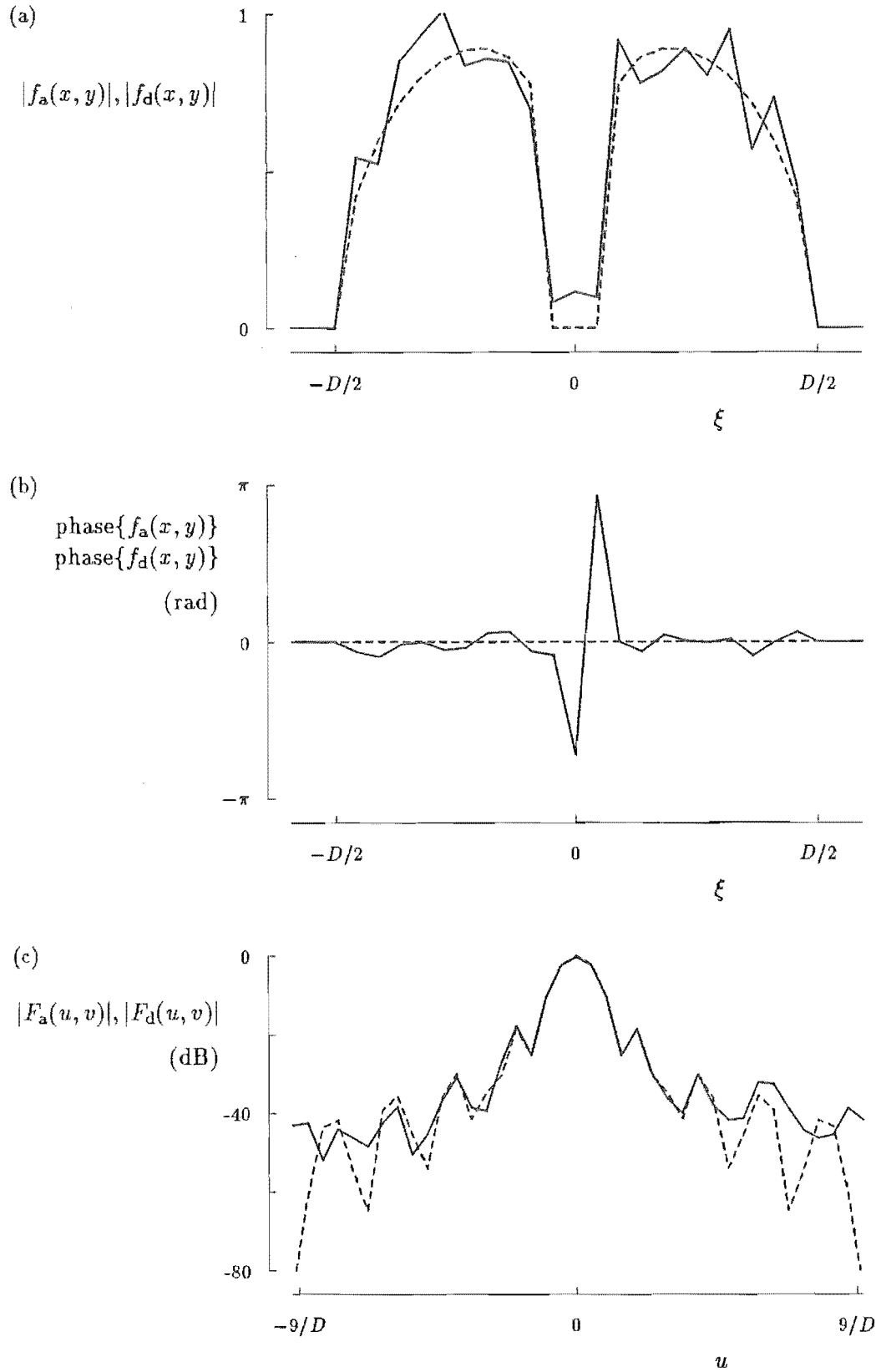


Figure 4.8 Actual copolar fields when $\tau_{\text{ran}} = 0.1$: (a) copolar aperture field amplitude distribution; (b) copolar aperture field phase distribution; (c) copolar far field amplitude distribution. The solid curves represent the actual copolar fields while the dashed curves represent the design copolar fields associated with design 2. The variable ξ, which identifies position along a diagonal in the aperture plane, is defined in (4.8).

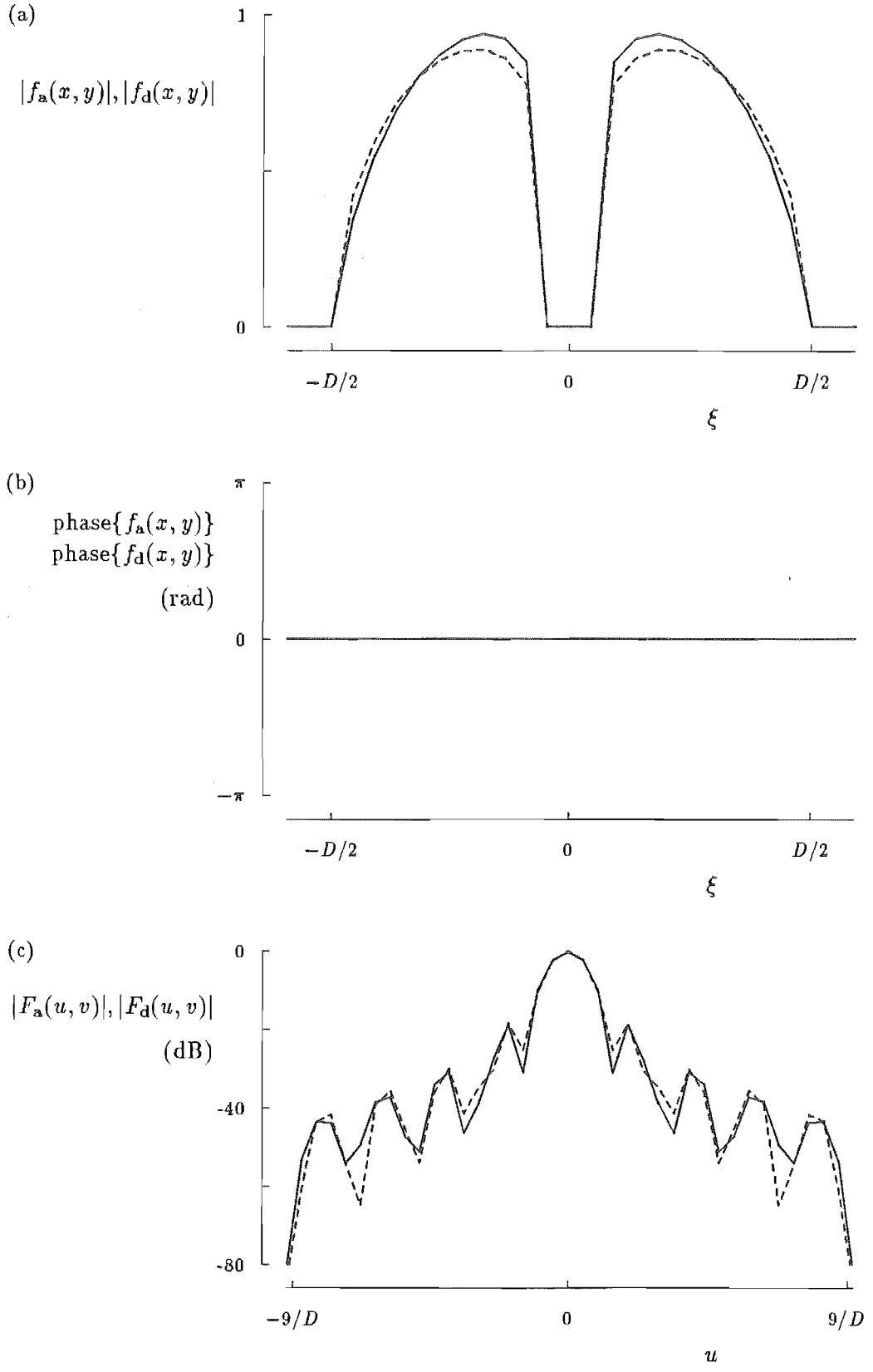


Figure 4.9 Actual copolar fields when $\tau_{\text{quad}} = 0.1$: (a) copolar aperture field amplitude distribution; (b) copolar aperture field phase distribution; (c) copolar far field amplitude distribution. The solid curves represent the actual copolar fields while the dashed curves represent the design copolar fields associated with design 2. The variable ξ, which identifies position along a diagonal in the aperture plane, is defined in (4.8).

4.2.3 Measurement inaccuracies

Far field *measurement inaccuracies* are here defined to be any differences between the actual copolar far field amplitude pattern $|F_{\mathrm{a}}(u,v)|$ and the measured copolar far field amplitude pattern $A_{\mathrm{m}}(u,v)$. In this section, three different causes of measurement inaccuracy, namely noise, calibration inaccuracy and data truncation, are modelled. The measured data are related to $|F_{\mathrm{a}}(u,v)|$ by

$$A_{\mathrm{m}}(u,v) = \begin{cases} \left| |F_{\mathrm{a}}(0,0)| \left(\dfrac{|F_{\mathrm{a}}(u,v)|}{|F_{\mathrm{a}}(0,0)|} \right)^{\Gamma_{\mathrm{cal}}} + \Gamma_{\mathrm{ran}} |F_{\mathrm{a}}(0,0)| \, \mathrm{ran}(u,v) \right| & \\ & \text{for } (u,v) \in S^{A_{\mathrm{m}}} \\ 0 & \text{elsewhere} \end{cases} \tag{4.15}$$

The quantities Γ_{cal}, Γ_{ran} and $S^{A_{\mathrm{m}}}$, are defined and discussed in the next three paragraphs respectively.

When measuring the amplitude pattern, it is convenient to employ an instrument which records the pattern in decibels relative to its peak value. In the absence of other measurement inaccuracies, $A_{\mathrm{m}}(u,v)$ and $|F_{\mathrm{a}}(u,v)|$ are then related by

$$20 \log_{10} \frac{A_{\mathrm{m}}(u,v)}{A_{\mathrm{m}}(0,0)} = \Gamma_{\mathrm{cal}} \, 20 \log_{10} \frac{|F_{\mathrm{a}}(u,v)|}{|F_{\mathrm{a}}(0,0)|} \tag{4.16}$$

where the left side of (4.16) represents the output of such an instrument and Γ_{cal} is a constant of proportionality which depends upon the calibration of the instrument. One method of calibrating the instrument involves using an attenuator in the following way. The peak signal is fed to the instrument which is adjusted so that it records 0 dB. The same signal is then attenuated before being fed to the instrument and the instrument is adjusted so that it records the level of attenuation expressed in decibels. This latter adjustment sets the value of Γ_{cal} to unity. It often transpires, of course, that the absolute level of the attenuation introduced by a given attenuator is not known exactly. This implies that the value of Γ_{cal} is not exactly 1.0 and furthermore cannot be accurately deduced. This type of calibration inaccuracy is modelled by the first term on the right side of (4.15). Comparison of this term with (4.16) reveals that the model sets $A_{\mathrm{m}}(0,0) = |F_{\mathrm{a}}(0,0)|$. The default value of Γ_{cal} is 1.0, which corresponds to the ideal case of a perfectly calibrated instrument. An example of calibration inaccuracy is illustrated in Figure 4.10(a).

No measurement can be performed with absolute precision, so there is always some uncertainty, or noise, associated with any measurement process. Morris [1985] points out that, for radio measurements, the dominant inaccuracy is likely to be additive receiver noise. This is modelled by a pseudo random distribution, represented by the second term on the right side of (4.15), added to the actual copolar far field amplitude pattern. The level of measurement noise is characterized by Γ_{ran}, which is the rms level of the noise relative to $|F_{\mathrm{a}}(0,0)|$. The default value of Γ_{ran} is 0. An example of the effect on $A_{\mathrm{m}}(u,v)$ of measurement noise is shown in Figure 4.10(b).

In the model, $S^{A_{\mathrm{m}}}$ is defined to be a disk of diameter $D^{A_{\mathrm{m}}}$ centred on the origin of the u,v plane. The default value of $D^{A_{\mathrm{m}}}$ is large enough for the disk to encompass the whole grid of far field sample points. When $D^{A_{\mathrm{m}}}$ is sufficiently small for some of the sample points to lie outside $S^{A_{\mathrm{m}}}$, the model simulates the situation in which the copolar far field amplitude pattern is not measured at these points. $A_{\mathrm{m}}(u,v)$ is therefore truncated because the sample values which are not measured are set to zero in (4.15).

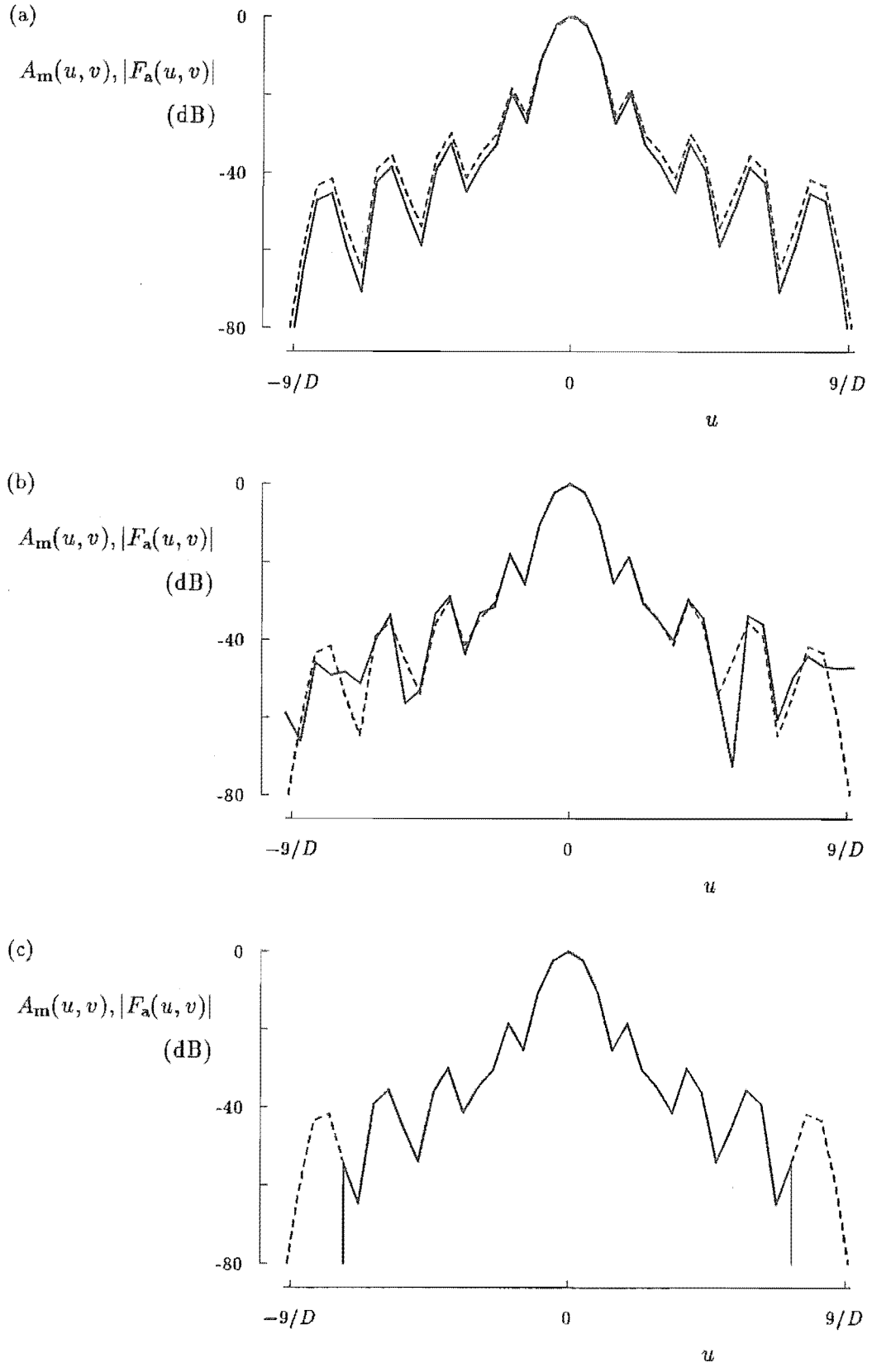


Figure 4.10 Measured copolar far field amplitude patterns for different measurement inaccuracies: (a) $\Gamma_{cal} = 1.1$; (b) $\Gamma_{ran} = -50$ dB; (c) $D^{A_m} = 15/D$. In all cases design 2 is employed. $A_m(u,v)$ and $|F_a(u,v)|$ are represented by a solid curve and a dashed curve respectively.

One motivation for truncating $A_{\rm m}(u,v)$ in this way is to simulate what happens when one attempts to reduce the duration and cost of a pattern measurement by reducing the number of measured samples of $A_{\rm m}(u,v)$. An example of $A_{\rm m}(u,v)$ suffering from truncation is illustrated in Figure 4.10(c).

4.2.4 Depolarization

The computer simulations presented in Section 4.9.1 refer to both copolar and cross polar fields. The purpose of this section is to describe extensions, to the model presented in Sections 4.2.1 to 4.2.3, which incorporate the cross polar fields. A linearly polarized antenna is modelled here. The cross polar field distributions are distinguished from the corresponding copolar field distributions by the superscript '×'.

The design copolar aperture field distribution $f_{\rm d}(x,y)$ is defined by either (4.6) or (4.7). The corresponding design cross polar aperture field distribution $f_{\rm d}^{\times}(x,y)$ is here defined by

$$f_{\rm d}^{\times}(x,y) = \tau_{\times\rm pol}\sin(2\phi)\, f_{\rm d}(x,y) \tag{4.17}$$

where $\tau_{\times\rm pol}$ characterizes the relative amplitudes of the cross polar and copolar aperture field distributions. The $\sin(2\phi)$ dependence in (4.17) is typical of the cross polar aperture field distribution produced by an axially symmetric antenna fed by almost any linearly polarized feed [Ghobrial, 1979; Rudge *et al.*, 1982, p. 397]. The design cross polar aperture field distribution is due to the intrinsic cross polar field associated with the feed.

As an intermediate step towards defining the actual aperture field distributions, the copolar and cross polar aperture field distributions $f_{\rm nt}(x,y)$ and $f_{\rm nt}^{\times}(x,y)$ are defined by

$$\begin{aligned} f_{\rm nt}(x,y) &= [|f_{\rm d}(x,y)| + \Delta a(x,y)]e^{j\,\Delta\psi(x,y)} \\ f_{\rm nt}^{\times}(x,y) &= \tau_{\times\rm pol}\sin(2\phi)\, f_{\rm nt}(x,y) \end{aligned} \tag{4.18}$$

where $\Delta a(x,y)$ and $\Delta\psi(x,y)$ are defined in Section 4.2.2. These distributions would be the actual aperture field distributions if the feed was not tilted (hence the subscript 'nt') and there was no scattering from struts. However, when these affects are present, the actual aperture field distributions are modelled by

$$\begin{aligned} f_{\rm a}(x,y) &= f_{\rm nt}(x,y) + \phi_{\rm tilt}\, f_{\rm nt}^{\times}(x,y) + n(x,y) \\ f_{\rm a}^{\times}(x,y) &= f_{\rm nt}^{\times}(x,y) - \phi_{\rm tilt}\, f_{\rm nt}(x,y) + n(x,y) \end{aligned} \tag{4.19}$$

where $n(x,y)$ models the field scattered by the struts and is defined by (4.14). It is assumed that the copolar and cross polar components of the scattered field are similar to each other. Provided it is small, the real number $\phi_{\rm tilt}$ represents the tilt angle (in radians) of the feed about the axis of the antenna. Therefore, $\phi_{\rm tilt}$ characterizes the angular difference between the design orientation and the actual orientation of the feed. The default value of $\phi_{\rm tilt}$ is zero radians.

It is assumed that the measuring apparatus does not introduce any further depolarization. Moreover, the only measurement inaccuracy that is modelled is measurement noise. Therefore, the measured copolar and cross polar far field amplitude patterns are modelled by

$$\begin{aligned} A_{\rm m}(u,v) &= ||F_{\rm a}(u,v)| + \Gamma_{\rm ran}\,|F_{\rm a}(0,0)|\,{\rm ran}(u,v)| \\ A_{\rm m}^{\times}(u,v) &= ||F_{\rm a}^{\times}(u,v)| + \Gamma_{\rm ran}\,|F_{\rm a}(0,0)|\,{\rm ran}(u,v)| \end{aligned} \tag{4.20}$$

Recall from Section 2.4.2.1 that an important design goal is to minimize the depolarization (as defined in Sec. 2.2.6) of the antenna. The depolarization of the simulated antenna, at the centre of its main beam, is defined to be $A_{\mathrm{m}}^{\mathrm{x}}(0,0)/A_{\mathrm{m}}(0,0)$. The existence of a cross polar field at the centre of the radiation pattern is due to several factors. While the cross polar field radiated by the feed, as modelled in (4.17), produces a cross polar far field pattern $F_{\mathrm{d}}^{\mathrm{x}}(u,v)$ which vanishes at its centre, any tilt of the feed adds a cross polar component at the pattern's centre. Geometrical defects tend to alter the phase of the cross polar aperture field distribution, thereby affecting $F_{\mathrm{a}}^{\mathrm{x}}(u,v)$, often increasing its amplitude at the centre of the pattern. So, a variety of geometrical defects plus any tilting of the feed degrade the cross polar far field pattern (as well as the copolar far field pattern).

4.3 ERROR MEASURES

As intimated in Section 4.1.1, the purpose of the modified Gerchberg-Saxton algorithm is to generate an estimate $f_{\mathrm{e}}(x,y)$ of the image-form of the actual copolar aperture field distribution $f_{\mathrm{a}}(x,y)$. The data to which the algorithm is applied are $A_{\mathrm{m}}(u,v)$ and $|f_{\mathrm{d}}(x,y)|$. In order to be able to assess how well the algorithm is performing in any particular instance, it is necessary to devise quantitative error measures indicating how satisfactory is the above-mentioned estimate $f_{\mathrm{e}}(x,y)$ generated by the algorithm.

It is useful to make a distinction between applying the modified Gerchberg-Saxton algorithm in practice and applying it in computer simulations. Should the algorithm be applied in practice, the only available information is represented by $A_{\mathrm{m}}(u,v)$ and $f_{\mathrm{d}}(x,y)$. Therefore, any error measures which are invoked in practice must not be predicated on knowledge of $f_{\mathrm{a}}(x,y)$. In computer simulations, however, the algorithm is applied to data generated from the computer model (Sec. 4.2), which also generates $f_{\mathrm{a}}(x,y)$. It makes sense, therefore, to utilize $f_{\mathrm{a}}(x,y)$ when formulating error measures for computer simulations. Four different error measures are introduced below. Two of them, denoted $\mathcal{E}^{\mathrm{ap}}$ and E_{c}, involve $f_{\mathrm{a}}(x,y)$ and so can only be invoked when the algorithm is applied in computer simulations. However, the other two error measures, denoted $\mathcal{E}^{\mathrm{fa}}$ and E_{m}, do not involve $f_{\mathrm{a}}(x,y)$ and could therefore be invoked in practice (as well as in computer simulations).

An indication of the accuracy of $\mathrm{phase}\{f_{\mathrm{e}}(x,y)\}$ is provided by the *aperture phase error* $\mathcal{E}^{\mathrm{ap}}$, which is defined to be the rms phase difference between $f_{\mathrm{e}}(x,y)e^{-j\psi^{\mathrm{ave}}}$ and the image-form of $f_{\mathrm{a}}(x,y)$, where ψ^{ave} is constant over $S^{f_{\mathrm{d}}}$. Because the addition of a constant phase to any field is physically irrelevant, ψ^{ave} is chosen to be the average phase difference between $f_{\mathrm{e}}(x,y)$ and the image-form of $f_{\mathrm{a}}(x,y)$, so that the value of $\mathcal{E}^{\mathrm{ap}}$ is minimized. Because the image-form of $f_{\mathrm{a}}(x,y)$ is ambiguous (Sec. 4.1.2), $\mathcal{E}^{\mathrm{ap}}$ is defined to be the rms difference between $\mathrm{phase}\{f_{\mathrm{e}}(x,y)e^{-j\psi^{\mathrm{ave}}}\}$ and whichever of $\mathrm{phase}\{f_{\mathrm{a}}(x,y)\}$ and $\mathrm{phase}\{\tilde{f}_{\mathrm{a}}(x,y)\}$ is closest to $\mathrm{phase}\{f_{\mathrm{e}}(x,y)e^{-j\psi^{\mathrm{ave}}}\}$. The computation of $\mathcal{E}^{\mathrm{ap}}$ is further complicated by the fact that phases can only be measured and/or computed modulo 2π. For this reason it is appropriate to introduce the notation $\langle\cdot\rangle_{\alpha}^{\alpha+2\pi}$, implying that whatever is in the angled brackets is expressed modulo 2π in the range α to $(\alpha+2\pi)$. $\mathcal{E}^{\mathrm{ap}}$ is taken to be the smallest of $\mathcal{E}_a^{\mathrm{ap}}$, $\mathcal{E}_b^{\mathrm{ap}}$, $\mathcal{E}_c^{\mathrm{ap}}$ and $\mathcal{E}_d^{\mathrm{ap}}$, where

$$
\begin{aligned}
\mathcal{E}_a^{\mathrm{ap}} &= \left[\frac{\iint_{S f_{\mathrm{d}}} \left[\left\langle \mathrm{phase}\{f_{\mathrm{e}}(x,y)\} - \mathrm{phase}\{f_{\mathrm{a}}(x,y)\} - \psi_a^{\mathrm{ave}}\right\rangle_0^{2\pi}\right]^2 dx\,dy}{\iint_{S f_{\mathrm{d}}} dx\,dy}\right]^{1/2} \\
\mathcal{E}_b^{\mathrm{ap}} &= \left[\frac{\iint_{S f_{\mathrm{d}}} \left[\left\langle \mathrm{phase}\{f_{\mathrm{e}}(x,y)\} - \mathrm{phase}\{f_{\mathrm{a}}(x,y)\} - \psi_b^{\mathrm{ave}}\right\rangle_{-\pi}^{\pi}\right]^2 dx\,dy}{\iint_{S f_{\mathrm{d}}} dx\,dy}\right]^{1/2} \\
\mathcal{E}_c^{\mathrm{ap}} &= \left[\frac{\iint_{S f_{\mathrm{d}}} \left[\left\langle \mathrm{phase}\{f_{\mathrm{e}}(x,y)\} - \mathrm{phase}\{\tilde{f}_{\mathrm{a}}(x,y)\} - \psi_c^{\mathrm{ave}}\right\rangle_0^{2\pi}\right]^2 dx\,dy}{\iint_{S f_{\mathrm{d}}} dx\,dy}\right]^{1/2} \\
\mathcal{E}_d^{\mathrm{ap}} &= \left[\frac{\iint_{S f_{\mathrm{d}}} \left[\left\langle \mathrm{phase}\{f_{\mathrm{e}}(x,y)\} - \mathrm{phase}\{\tilde{f}_{\mathrm{a}}(x,y)\} - \psi_d^{\mathrm{ave}}\right\rangle_{-\pi}^{\pi}\right]^2 dx\,dy}{\iint_{S f_{\mathrm{d}}} dx\,dy}\right]^{1/2}
\end{aligned}
\tag{4.21}
$$

with ψ_a^{ave}, ψ_b^{ave}, ψ_c^{ave} and ψ_d^{ave} being respectively defined by

$$
\begin{aligned}
\psi_a^{\mathrm{ave}} &= \frac{\iint_{S f_{\mathrm{d}}} \left\langle \mathrm{phase}\{f_{\mathrm{e}}(x,y)\} - \mathrm{phase}\{f_{\mathrm{a}}(x,y)\}\right\rangle_0^{2\pi} dx\,dy}{\iint_{S f_{\mathrm{d}}} dx\,dy} \\
\psi_b^{\mathrm{ave}} &= \frac{\iint_{S f_{\mathrm{d}}} \left\langle \mathrm{phase}\{f_{\mathrm{e}}(x,y)\} - \mathrm{phase}\{f_{\mathrm{a}}(x,y)\}\right\rangle_{-\pi}^{\pi} dx\,dy}{\iint_{S f_{\mathrm{d}}} dx\,dy} \\
\psi_c^{\mathrm{ave}} &= \frac{\iint_{S f_{\mathrm{d}}} \left\langle \mathrm{phase}\{f_{\mathrm{e}}(x,y)\} - \mathrm{phase}\{\tilde{f}_{\mathrm{a}}(x,y)\}\right\rangle_0^{2\pi} dx\,dy}{\iint_{S f_{\mathrm{d}}} dx\,dy} \\
\psi_d^{\mathrm{ave}} &= \frac{\iint_{S f_{\mathrm{d}}} \left\langle \mathrm{phase}\{f_{\mathrm{e}}(x,y)\} - \mathrm{phase}\{\tilde{f}_{\mathrm{a}}(x,y)\}\right\rangle_{-\pi}^{\pi} dx\,dy}{\iint_{S f_{\mathrm{d}}} dx\,dy}
\end{aligned}
\tag{4.22}
$$

The value of ψ^{ave} is either ψ_a^{ave}, or ψ_b^{ave}, or ψ_c^{ave} or ψ_d^{ave}, when $\mathcal{E}^{\mathrm{ap}}$ is taken to be either $\mathcal{E}_a^{\mathrm{ap}}$, or $\mathcal{E}_b^{\mathrm{ap}}$, or $\mathcal{E}_c^{\mathrm{ap}}$ or $\mathcal{E}_a^{\mathrm{ap}}$ respectively. Note that the only operations in (4.21) and (4.22) which are computed with modulo arithmetic are those represented by angled brackets. The rms phase difference between $f_{\mathrm{e}}(x,y)e^{-j\psi^{\mathrm{ave}}}$ and $f_{\mathrm{a}}(x,y)$ is the smaller of $\mathcal{E}_a^{\mathrm{ap}}$ and $\mathcal{E}_b^{\mathrm{ap}}$, while the rms phase difference between $f_{\mathrm{e}}(x,y)e^{-j\psi^{\mathrm{ave}}}$ and $\tilde{f}_{\mathrm{a}}(x,y)$ is the smaller of $\mathcal{E}_c^{\mathrm{ap}}$ and $\mathcal{E}_d^{\mathrm{ap}}$. To understand why both $\mathcal{E}_a^{\mathrm{ap}}$ and $\mathcal{E}_b^{\mathrm{ap}}$ must be computed, consider the following three situations.

In the first of these situations, $f_{\mathrm{e}}(x,y)e^{-j\pi}$ and $f_{\mathrm{a}}(x,y)$ are approximately equal, so that $\langle \mathrm{phase}\{f_{\mathrm{e}}(x,y)\} - \mathrm{phase}\{f_{\mathrm{a}}(x,y)\}\rangle_0^{2\pi}$ is close to π rad. However, $\langle \mathrm{phase}\{f_{\mathrm{e}}(x,y)\} - \mathrm{phase}\{f_{\mathrm{a}}(x,y)\}\rangle_{-\pi}^{\pi}$ is slightly greater than $-\pi$ rad. at some points (x,y), but is slightly less than π rad. at other points (x,y). Therefore, it follows from the first and second equations of (4.22) that $\psi_a^{\mathrm{ave}} = \pi$ rad., which is the correct average phase difference between $f_{\mathrm{e}}(x,y)$ and $f_{\mathrm{a}}(x,y)$, but $\psi_b^{\mathrm{ave}} \approx 0$ rad. Similar reasoning shows that while

$\mathcal{E}_a^{\mathrm{ap}}$ is small and correctly indicates the rms phase difference between $f_{\mathrm{e}}(x,y)e^{-j\pi}$ and $f_{\mathrm{a}}(x,y)$, $\mathcal{E}_b^{\mathrm{ap}}$ is always much larger.

In the second situation, $f_{\mathrm{e}}(x,y)$ is approximately equal to $f_{\mathrm{a}}(x,y)$. Similar reasoning to that invoked in the previous paragraph shows that $\mathcal{E}_b^{\mathrm{ap}}$ is the correct rms phase difference between $f_{\mathrm{e}}(x,y)$ and $f_{\mathrm{a}}(x,y)$, while $\mathcal{E}_a^{\mathrm{ap}}$ is always much larger.

In the third situation, which is more general than either of the situations considered in the previous two paragraphs, $f_{\mathrm{e}}(x,y)e^{-j\psi^{\mathrm{ave}}}$ and $f_{\mathrm{a}}(x,y)$ are approximately equal, where ψ^{ave} is arbitrary. The value of ψ^{ave} cannot be estimated before (4.22) and (4.21) are evaluated. Therefore, both $\mathcal{E}_a^{\mathrm{ap}}$ and $\mathcal{E}_b^{\mathrm{ap}}$ must be calculated, after which the smaller of the two can be immediately recognized.

The *far field error* $\mathcal{E}^{\mathrm{fa}}$ is a measure of how close are $|F_{\mathrm{e}}(u,v)|$ and $A_{\mathrm{m}}(u,v)$. It is defined by

$$\mathcal{E}^{\mathrm{fa}} = \frac{1}{A_{\mathrm{m}}(0,0)} \left[\frac{\iint \left[|F_{\mathrm{e}}(u,v)| - A_{\mathrm{m}}(u,v) \right]^2 \, du\, dv}{\iint du\, dv} \right]^{1/2} \tag{4.23}$$

Recall that, as specified in Section 3.4.3, both integrals in (4.23) are performed over the region of the u,v plane spanned by the grid of sample points. In words, $\mathcal{E}^{\mathrm{fa}}$ is the rms difference between $|F_{\mathrm{e}}(u,v)|$ and $A_{\mathrm{m}}(u,v)$ normalized with respect to the peak value of $A_{\mathrm{m}}(u,v)$, which is taken for convenience to be at the centre of the radiation pattern. The normalization is similar to that implicit in Γ_{ran} (Sec. 4.2.3), which characterizes the amount of far field measurement noise in the computer model. Because of this, values of $\mathcal{E}^{\mathrm{fa}}$ and of Γ_{ran} can be compared directly with each other. Note that $\mathcal{E}^{\mathrm{fa}}$ is similar to the Fourier error $\mathcal{E}^{\mathrm{F}}$, which is discussed in Section 3.4.3.1 in the context of the original Gerchberg-Saxton algorithm.

The purpose of generating $f_{\mathrm{e}}(x,y)$ is to help correct any geometrical defects of an antenna so that the radiation pattern can meet its specifications. To establish whether the corrected radiation pattern can be expected to be an improvement over the original radiation pattern, it is desirable to define an error which indicates how well either of these radiation patterns meets its specifications. As intimated in Section 2.4.2.2, the specifications are often in the form of an envelope under which the radiation pattern levels must lie. In the following paragraph a model of an envelope is described so that an envelope error can be defined.

The *design envelope* $\Lambda_{\mathrm{d}}(u,v)$ is a smooth, real-valued function which is never less than the design copolar far field amplitude pattern $|F_{\mathrm{d}}(u,v)|$ plus an offset $\Gamma_{\mathrm{off}}\,|F_{\mathrm{d}}(0,0)|$. The purpose of the envelope offset is discussed later in this section. Because the models of the design fields (Sec. 4.2.1) are circularly symmetric, $\Lambda_{\mathrm{d}}(u,v)$ is also chosen to be circularly symmetric. The design envelope is equal to $[|F_{\mathrm{d}}(u,v)| + \Gamma_{\mathrm{off}}\,|F_{\mathrm{d}}(0,0)|]$ over the central portion of the main beam and at the peaks of all the sidelobes of the radiation pattern. Overall, $\Lambda_{\mathrm{d}}(u,v)$ decreases monotonically away from the centre of the radiation pattern. Figure 4.11 graphs a cross-section through $\Lambda_{\mathrm{d}}(u,v)$ and $|F_{\mathrm{d}}(u,v)|$ for design 1 with $\Gamma_{\mathrm{off}}\,(u,v) = 0.002$.

The *measured envelope error* E_{m} is a measure of the maximum amount by which the measured copolar far field amplitude pattern $A_{\mathrm{m}}(u,v)$ exceeds the design envelope $\Lambda_{\mathrm{d}}(u,v)$. It is defined by

$$E_{\mathrm{m}} = \max \left(20log \frac{A_{\mathrm{m}}(u,v)}{A_{\mathrm{m}}(0,0)} - 20log \frac{\Lambda_{\mathrm{d}}(u,v)}{\Lambda_{\mathrm{d}}(0,0)} \right) \tag{4.24}$$

where $\max(\cdot)$ denotes selecting the maximum value. In words, E_{m} is the maximum amount by which $A_{\mathrm{m}}(u,v)$ exceeds $\Lambda_{\mathrm{d}}(u,v)$ after both have been expressed in decibels

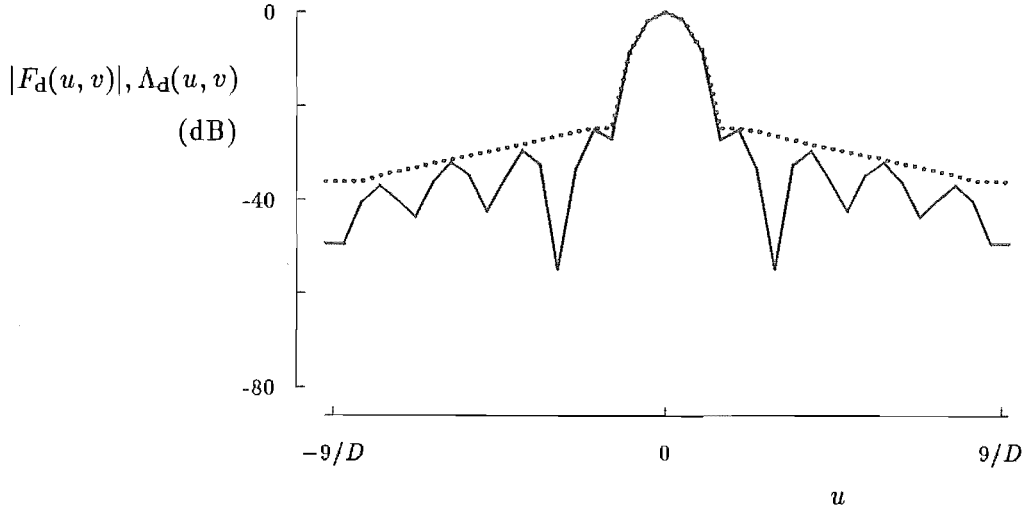


Figure 4.11 Design envelope for design 1. A cross-section through the design envelope $\Lambda_d(u,v)$ (dotted curve) is compared with the corresponding cross-section through $|F_d(u,v)|$ (solid curve) where $\Gamma_{off} = 0.002$.

relative to their respective peak values. In the following paragraph it is assumed that $\Gamma_{off} = 0$.

The measured envelope error E_m quantitates the following three undesirable characteristics (as intimated in Sec. 2.5) of $A_m(u,v)$: the absolute sidelobe levels of $A_m(u,v)$ exceeding those of $|F_d(u,v)|$, the absolute peak level of $A_m(u,v)$ exceeding the absolute peak level of $|F_d(u,v)|$, and the main beam of $A_m(u,v)$ being wider than that of $|F_d(u,v)|$. Either of the first two of these characteristics results in the sidelobe levels of $A_m(u,v)/A_m(0,0)$ being higher than those of $|F_d(u,v)|/|F_d(0,0)|$. This in turn implies that the sidelobe levels of $A_m(u,v)/A_m(0,0)$ exceed $\Lambda_d(u,v)/\Lambda_d(0,0)$ because $\Lambda_d(u,v)$ equals $|F_d(u,v)|$ at the sidelobe peaks of $|F_d(u,v)|$. When the beamwidth of $A_m(u,v)$ is wider than that of $|F_d(u,v)|$, $A_m(u,v)/A_m(0,0)$ must be greater than $|F_d(u,v)|/|F_d(0,0)|$ over the region of the u,v plane corresponding to the main beam of $|F_d(u,v)|$, but excluding the point $(u,v) = (0,0)$. Therefore, because $\Lambda_d(u,v)$ is equal to $|F_d(u,v)|$ over the central portion of the main beam, $A_m(u,v)/A_m(0,0)$ must be greater than $\Lambda_d(u,v)/\Lambda_d(0,0)$ over this region of the u,v plane, but excluding the point $(0,0)$. Since E_m is a measure of the maximum amount by which $A_m(u,v)/A_m(0,0)$ exceeds $\Lambda_d(u,v)/\Lambda_d(0,0)$, its value is affected by any of the undesirable characteristics of $A_m(u,v)$ which are mentioned in this paragraph.

An undesirable aspect of the above definition of E_m is that it can be unduly affected by measurement inaccuracies. Consider an antenna which is behaving ideally so that $|F_a(u,v)| = |F_d(u,v)|$. Suppose, first, that $|F_a(u,v)|$ is measured with infinite accuracy, so that $A_m(u,v) = |F_a(u,v)| = |F_d(u,v)|$. This implies that $E_m = 0$ dB. However, when the inevitable measurement noise, which is characterized by Γ_{ran} (Sec. 4.2.3), is included, one or more of the sidelobe peaks of $A_m(u,v)/A_m(0,0)$ is likely to exceed $\Lambda_d(u,v)/\Lambda_d(0,0)$, thereby causing E_m to exceed 0 dB. E_m can be reduced by setting the envelope offset Γ_{off} to a positive value. Recalling the definitions, given earlier in this section, of Γ_{off} and $\Lambda_d(u,v)$, it is seen that, when $\Gamma_{off} = 2\,\Gamma_{ran}$, the amount by which $\Lambda_d(u,v)$ exceeds the sidelobe peaks of $|F_d(u,v)|$ is about the same as the maximum

amount by which the sidelobe peaks of $A_{\mathrm{m}}(u,v)$ can exceed the sidelobe peaks of $|F_{\mathrm{a}}(u,v)|$. This implies that, when $\Gamma_{\mathrm{off}} = 2\,\Gamma_{\mathrm{ran}}$, measurement noise is unlikely to cause $A_{\mathrm{m}}(u,v)/A_{\mathrm{m}}(0,0)$ to exceed $\Lambda_{\mathrm{d}}(u,v)/\Lambda_{\mathrm{d}}(0,0)$. Therefore, provided Γ_{off} is set to an appropriate value, E_{m} is unlikely to be affected by measurement noise. Note, however, that E_{m} is affected by calibration inaccuracy in the same way that it is affected by measurement noise when $\Gamma_{\mathrm{off}} = 0$.

A measure similar to E_{m} can indicate how well the corrected radiation pattern meets its specifications. Before defining this new measure, a way of modelling the correction process must first be introduced. As outlined in Section 4.1.1, the information contained in $f_{\mathrm{e}}(x,y)$ is utilized to determine, and then remove, the geometrical defects of the antenna. This results in what is here called the *corrected copolar aperture field distribution* $f_{\mathrm{c}}(x,y)$. Assuming the geometrical defects to be small, correcting them affects the amplitude of the copolar aperture field distribution negligibly for reasons given in Section 3.2.1. Provided that $f_{\mathrm{e}}(x,y)$ is a good approximation to $f_{\mathrm{a}}(x,y)$, the geometrical corrections alter the phase of the copolar aperture field distribution by $(-\,\mathrm{phase}\{f_{\mathrm{e}}(x,y)\} + \psi_0)$, thereby forcing $\mathrm{phase}\{f_{\mathrm{c}}(x,y)\}$ to be almost uniform over $S^{f_{\mathrm{d}}}$. The real number ψ_0, which is arbitrary because its value does not affect $|F_{\mathrm{c}}(u,v)|$, can be conveniently chosen to minimize the above-mentioned geometrical corrections. Because of the above reasoning, $f_{\mathrm{c}}(x,y)$ is here modelled by

$$f_{\mathrm{c}}(x,y) = \begin{cases} f_{\mathrm{a}}(x,y)e^{-j[\mathrm{phase}\{f_{\mathrm{e}}(x,y)\}-\psi_0]} & \text{if } f_{\mathrm{e}}(x,y) \approx f_{\mathrm{a}}(x,y) \\ f_{\mathrm{a}}(x,y)e^{j[\mathrm{phase}\{f_{\mathrm{e}}(-x,-y)\}+\psi_0]} & \text{if } f_{\mathrm{e}}(x,y) \approx \tilde{f}_{\mathrm{a}}(x,y) \end{cases} \tag{4.25}$$

Section 4.1.2 suggests how, in practice, it can be determined whether $f_{\mathrm{a}}(x,y)$ or $\tilde{f}_{\mathrm{a}}(x,y)$ is approximated by $f_{\mathrm{e}}(x,y)$. In computer simulations, the ambiguity can be resolved by computing $\mathcal{E}^{\mathrm{ap}}$ in the way described earlier in this section. If $\mathcal{E}^{\mathrm{ap}}$ equals either $\mathcal{E}^{\mathrm{ap}}_a$ or $\mathcal{E}^{\mathrm{ap}}_b$, $f_{\mathrm{e}}(x,y)$ is deemed to approximate $f_{\mathrm{a}}(x,y)$. Otherwise, $f_{\mathrm{e}}(x,y)$ approximates $\tilde{f}_{\mathrm{a}}(x,y)$. ψ_0 is taken to equal ψ^{ave} so that, as follows from the argument developed earlier in this section, the average value of $\mathrm{phase}\{f_{\mathrm{c}}(x,y)\}$ is zero.

The *corrected envelope error* E_{c}, which is a measure of how well $F_{\mathrm{c}}(u,v)$ meets its specifications, is defined by (cf. (4.24))

$$E_{\mathrm{c}} = \max\left(20log\frac{|F_{\mathrm{c}}(u,v)|}{|F_{\mathrm{c}}(0,0)|} - 20log\frac{\Lambda_{\mathrm{d}}(u,v)}{\Lambda_{\mathrm{d}}(0,0)}\right) \tag{4.26}$$

Comparison of (4.26) with (4.24) indicates that the corrected and measured envelope errors are similar to each other. Therefore, much of the discussion, earlier in this section, about E_{m} is also relevant to E_{c}. However, an advantage of E_{c} over E_{m} is that the former is not directly affected by any measurement inaccuracies, because $|F_{\mathrm{c}}(u,v)|$ is not a measured pattern.

The errors $\mathcal{E}^{\mathrm{fa}}$, $\mathcal{E}^{\mathrm{ap}}$, E_{m} and E_{c} are now compared with each other, from the point of view of how well they indicate the usefulness of the estimate $f_{\mathrm{e}}(x,y)$ generated by a particular run of the modified Gerchberg-Saxton algorithm.

Should the modified Gerchberg-Saxton algorithm be applied in practice, $\mathcal{E}^{\mathrm{fa}}$ is the only one of the errors defined in this section which can be computed to provide a measure of the performance of the algorithm: calculation of $\mathcal{E}^{\mathrm{ap}}$ and E_{c} both require knowledge of $f_{\mathrm{a}}(x,y)$, while E_{m} is not a measure of the accuracy of $f_{\mathrm{e}}(x,y)$. However, a disadvantage of $\mathcal{E}^{\mathrm{fa}}$ is that it is only an indirect measure of the accuracy of $f_{\mathrm{e}}(x,y)$. This is because it provides a measure of how accurately $|F_{\mathrm{e}}(u,v)|$ estimates $A_{\mathrm{m}}(u,v)$, which is only an approximation to $|F_{\mathrm{a}}(u,v)|$.

In computer simulations, a more direct measure of the accuracy of $f_e(x,y)$ is provided by $\mathcal{E}^{ap}$. $f_e(x,y)$ can be judged to be sufficiently accurate if $\mathcal{E}^{ap}$ falls below a preset threshold dependent upon the measurement inaccuracies. Note that the value of $\mathcal{E}^{ap}$ also provides an indication of the accuracy to which the shape defects of the main reflector can be computed. This can be understood by considering an antenna in which the aperture phase deviations are entirely due to shape defects of the main reflector of the antenna. For a shallow reflector, the shape defects $\Delta n(x,y)$ are related to aperture phase deviations by (3.3). Because aperture phase deviations are the phase differences between $f_a(x,y)$ and $f_d(x,y)$, (3.3) can be rearranged to yield

$$\Delta n(x,y) = \frac{1}{2k}[\text{phase}\{f_a(x,y)\} - \text{phase}\{f_d(x,y)\}] \tag{4.27}$$

However, should the modified Gerchberg-Saxton algorithm be used in practice, $f_a(x,y)$ must be replaced by its estimate, which is either $f_e(x,y)e^{-j\psi^{ave}}$ or $\tilde{f}_e(x,y)e^{-j\psi^{ave}}$. Assuming that $f_e(x,y)e^{-j\psi^{ave}}$ is the estimate of $f_a(x,y)$, an estimate $\Delta n_e(x,y)$ of the shape defects is

$$\Delta n_e(x,y) = \frac{1}{2k}[\text{phase}\{f_e(x,y)\} - \psi^{ave} - \text{phase}\{f_d(x,y)\}] \tag{4.28}$$

The accuracy to which the shape defects are known is greater the smaller the difference between $\Delta n_e(x,y)$ and $\Delta n(x,y)$, an expression for which can be obtained by eliminating $f_d(x,y)$ from (4.27) and (4.28):

$$\Delta n_e(x,y) - \Delta n(x,y) = \frac{1}{2k}[\text{phase}\{f_e(x,y)\} - \text{phase}\{f_a(x,y)\} - \psi^{ave}] \tag{4.29}$$

The rms values of the two sides of (4.29) must equal each other. The rms value of the term in square brackets in (4.29) is, by definition, $\mathcal{E}^{ap}$. The rms accuracy to which the geometrical defects can be estimated is therefore $\mathcal{E}^{ap}/2k$. Note, however, that when there is significant scattering from struts, the aperture phase deviation is due in part to geometrical defects and in part to the scattered field. Therefore, the accuracy of the estimated shape defects, calculated from (4.28), is limited by the amplitude (characterized, in the computer model, by τ_{ran} in (4.14)) of the scattered field [Kerbyson *et al.*, 1987].

The way in which E_m and E_c provide an indication of whether or not $A_m(u,v)$ and $|F_c(u,v)|$, respectively, meet their specifications is now described. The maximum sidelobe levels allowed by the specifications are here denoted by $\Lambda_s(u,v)$. For the purposes of the discussion in this paragraph, the quantities $\Lambda_s(u,v)$, $\Lambda_d(u,v)$, $|F_d(u,v)|$, $|F_a(u,v)|$ and $A_m(u,v)$ are taken to be expressed in decibels. An antenna is typically designed so that the peak sidelobe levels of $|F_d(u,v)|$ are less than $\Lambda_s(u,v)$ by at least a *design safety margin* E_{sm}. It follows that, assuming $\Gamma_{off} = 0$, $\Lambda_d(u,v)$ must also be less than $\Lambda_s(u,v)$, by at least E_{sm}, over the region of the u,v plane which excludes the main beam of $|F_d(u,v)|$. Because E_m is a measure of the maximum amount by which $A_m(u,v)$ exceeds $\Lambda_d(u,v)$, it follows that, provided $E_m \leq E_{sm}$, $A_m(u,v)$ does not exceed $\Lambda_s(u,v)$. Therefore, by definition, $A_m(u,v)$ meets its specifications if $E_m \leq E_{sm}$. For the purposes of this thesis, $A_m(u,v)$ is deemed to have failed to meet its specifications if $E_m > E_{sm}$. Similarly, $|F_c(u,v)|$ is deemed to have met its specifications only if $E_c \leq E_{sm}$.

Of the measures described in this section, E_c is conceptually the best measure of the accuracy of phase$\{f_e(x,y)\}$. This is because it assesses the end result of the whole

process, described in Section 4.1.1, involving both the generation of $f_e(x,y)$ and the correcting of the antenna. Even when $\mathcal{E}^{ap}$ is judged to be too large, provided $E_c \leq E_{sm}$, $f_e(x,y)$ is a usefully accurate estimate of the image-form of $f_a(x,y)$. Ideally, E_c should be zero, which would imply that the corrections to the antenna's geometry have restored the antenna to an equivalent of its design state.

It is important to realize that the value of E_c is not only affected by the accuracy of phase$\{f_e(x,y)\}$, generated by the modified Gerchberg-Saxton algorithm, but is also affected by any differences between $|f_c(x,y)|$ and $|f_d(x,y)|$. From (4.25) is it apparent that

$$|f_c(u,v)| - |f_d(u,v)| = |f_a(u,v)| - |f_d(u,v)| \tag{4.30}$$

This is because the correction process, which is modelled by (4.25), only incorporates those changes to the geometry of the antenna which attempt to set the phase differences between $f_c(x,y)$ and $f_d(x,y)$ to zero. However, no changes are made to the copolar amplitude distribution of the aperture field. Therefore, it is possible for the differences (characterized in the computer model by τ_{quad} and τ_{ran}) between $|f_c(x,y)|$ and $|f_d(x,y)|$ to be so great that $E_c > E_{sm}$, no matter how accurately $f_e(x,y)$ approximates $f_a(x,y)$.

In Section 4.4.5, it is suggested that different forms of the modified Gerchberg-Saxton algorithm should be applied to the same data to produce several different estimates $f_e(x,y)$, the most accurate of which is selected. Because it is felt that this is a procedure which could be implemented in practice, the accuracy of each estimate must be determined on the basis of $\mathcal{E}^{fa}$. Accordingly, the 'best' estimate is always selected to be the one for which $\mathcal{E}^{fa}$ is smallest. For the computer simulations of this procedure presented in this chapter, the accuracy of the selected $f_e(x,y)$ is determined by the corresponding values of $\mathcal{E}^{ap}$, E_m and E_c. Whereas $\mathcal{E}^{ap}$ is the most direct measure of the accuracy of phase$\{f_e(x,y)\}$, E_c is perhaps more appropriate, for reasons given earlier in this section. The values of E_c and E_m can be compared with each other to reveal the improvement (or otherwise) of $|F_c(u,v)|$ over $A_m(u,v)$.

4.4 THE MODIFIED GERCHBERG-SAXTON ALGORITHM

The *modified Gerchberg-Saxton algorithm* is a generalization of the original Gerchberg-Saxton algorithm, described in Section 3.4.3.1. It is also a specialization of the basic iterative Fourier transform algorithm, which is introduced in Section 3.4.3 and is illustrated in Figure 3.9. The information needed for the original Gerchberg-Saxton algorithm is $[f(x,y)] = |f(x,y)|$ and $[F(u,v)] = |F(u,v)|$. In applications for which the modified Gerchberg-Saxton algorithm is appropriate, $|F_a(u,v)|$ is measured, so that $[F(u,v)] = A_m(u,v)$. However, in these applications, $|F_a(x,y)|$ is not measured, but $|f_d(x,y)|$ is available (see Sec. 4.1.1), so it is appropriate to set $[f(x,y)] = |f_d(x,y)|$. It must not be forgotten that inherent in $[f(x,y)]$ is information about the aperture support S^{aper}. The main difference between the original and modified Gerchberg-Saxton algorithms lies in the way the constraints are applied. The original Gerchberg-Saxton algorithm utilizes an error reduction type of constraint which is defined by (3.59). A serious disadvantage of the original Gerchberg-Saxton algorithm is that it usually stagnates far from the correct solution. In an effort to avoid this, a variety of constraints can be incorporated into the modified Gerchberg-Saxton algorithm. Thus, there are many different possible forms of the modified Gerchberg-Saxton algorithm, which differ from each other according to whatever particular constraints are invoked. The more successful of these forms are employed in the computer simulations presented in this

chapter and are described in Sections 4.4.1 to 4.4.5.

Section 4.4.1 describes, in terms of the modified Gerchberg-Saxton algorithm, the original Gerchberg-Saxton algorithm and Fienup's error reduction algorithm. Forms of the modified Gerchberg-Saxton algorithm which are based on Gerchberg's variant on the Gerchberg-Saxton algorithm and on Fienup's hybrid input-output algorithm are discussed in Sections 4.4.2 and 4.4.3 respectively. The question of what should be the starting aperture distribution for the modified Gerchberg-Saxton algorithm is addressed in Section 4.4.4. Because the different forms of the modified Gerchberg-Saxton algorithm are appropriate in different situations, a composite algorithm is introduced in Section 4.4.5. This comprises several runs of different forms of the algorithm followed by a decision as to which of these runs produced the best solution. Other forms of the modified Gerchberg-Saxton algorithm, which I have examined, are discussed in Section 4.4.6.

Note that the errors $\mathcal{E}^{\mathrm{ap}}$ and $\mathcal{E}^{\mathrm{fa}}$, defined by (4.21) and (4.23) respectively, can be computed at each iteration of the algorithm. This is affected in the i^{th} iteration by setting $f_{\mathrm{e}}(x,y) = g_i(x,y)$ and equivalently setting $F_{\mathrm{e}}(u,v) = G_i(x,y)$, where $g_i(x,y)$ is the i^{th} estimate, generated by the modified Gerchberg-Saxton algorithm, of the copolar aperture field distribution. For the computer simulations presented in Sections 4.4.1 to 4.4.4, $\mathcal{E}_i^{\mathrm{ap}}$ is invoked in preference to $\mathcal{E}_i^{\mathrm{fa}}$, because $\mathcal{E}_i^{\mathrm{ap}}$ is the better indicator of the degree of convergence of the algorithm (Sec. 4.3). An *error curve* is here defined to be a graph of $\mathcal{E}_i^{\mathrm{ap}}$ versus i. For each error curve, it is useful to determine a *target value (error measure)* $\mathcal{E}_{\mathrm{targ}}^{\mathrm{ap}}$ for $\mathcal{E}_i^{\mathrm{ap}}$. $\mathcal{E}_{\mathrm{targ}}^{\mathrm{ap}}$ is defined, before the modified Gerchberg-Saxton algorithm is applied to particular data, to be the smallest value of $\mathcal{E}_i^{\mathrm{ap}}$ that can be expected to be achieved, given the quality of those data. Any particular run of the modified Gerchberg-Saxton algorithm applied to those particular data can then be said to converge satisfactorily if $\mathcal{E}_i^{\mathrm{ap}}$ falls to a value less than or close to $\mathcal{E}_{\mathrm{targ}}^{\mathrm{ap}}$. The parameters of the particular models invoked in Sections 4.4.1 to 4.4.4 have values similar to those of the parameters describing the models invoked to generate the results shown in Figure 4.37. It is my experience that the particular modified Gerchberg-Saxton algorithm runs, which generated the results shown in Figure 4.37, converged as well as can be expected, given the model parameters involved. For convenience, therefore, the values of $\mathcal{E}_{\mathrm{targ}}^{\mathrm{ap}}$, invoked in Sections 4.4.1 to 4.4.4, are taken from the results displayed in Figure 4.37 when the modified Gerchberg-Saxton algorithm is applied to data generated from models for which $\mathcal{E}^{\mathrm{fa}} = -60$ dB and $\mathcal{E}^{\mathrm{fa}} = -70$ dB respectively. The corresponding values of $\mathcal{E}_{\mathrm{targ}}^{\mathrm{ap}}$ are 0.033 and 0.010 radians. The value of $\mathcal{E}_{\mathrm{targ}}^{\mathrm{ap}}$ applicable to each particular run of the modified Gerchberg-Saxton algorithm is indicated by an arrow on the corresponding error curve.

4.4.1 Error reduction algorithms

When expressed in the radio engineering notation introduced in Section 4.2, one iteration of Fienup's error reduction algorithm for complex images (Sec. 3.4.3.3) is described by (cf. (3.70))

$$\begin{aligned} G_i(u,v) &= \mathrm{FT}\{g_i(x,y)\} \\ G_i'(u,v) &= A_{\mathrm{m}}(u,v)e^{j\,\mathrm{phase}\{G_i(u,v)\}} \\ g_i'(x,y) &= \mathrm{IFT}\{G_i'(u,v)\} \\ g_{i+1}(x,y) &= \begin{cases} g_i'(x,y) & \text{for } (x,y) \in S^{\mathrm{aper}} \\ 0 & \text{elsewhere} \end{cases} \end{aligned} \tag{4.31}$$

Figure 4.12(a) depicts an error curve for one thousand iterations of Fienup's error reduction algorithm applied to the particular model defined by

$$\begin{aligned} &\text{design 2,} \quad \psi_{\mathrm{quad}} = 2.0 \text{ rad.}; \quad \tau_{\mathrm{ran}} = 0.01; \\ &\Gamma_{\mathrm{ran}} = 0.001 = -60 \text{ dB}; \qquad \Gamma_{\mathrm{off}} = 0.002 \end{aligned} \tag{4.32}$$

Recall that the design and all of the parameters, listed in (4.32), are defined in Section 4.2 and that the model parameters not specified in the definition of any particular model assume their default values. The starting aperture distribution utilized for Fienup's error reduction algorithm is

$$g_1(x,y) = |f_{\mathrm{d}}(x,y)|e^{j\pi\,\mathrm{ran}(x,y)/\sqrt{3}} \tag{4.33}$$

where the exponential term is chosen so that the values of $\mathrm{phase}\{g_1(x,y)\}$ vary randomly from sample to sample and are uniformly distributed between $-\pi$ and π. Unless otherwise specified, all runs of the modified Gerchberg-Saxton algorithm discussed in this thesis are started with an aperture distribution having the form expressed by (4.33). The three different curves in Figure 4.12(a) correspond to running Fienup's error reduction algorithm starting with three different random starting phase distributions for $g_1(x,y)$.

It is clear from Figure 4.12(a) that all three of the runs of Fienup's error reduction algorithm have stagnated for values of $\mathcal{E}^{\mathrm{ap}}_{1000}$ much greater than $\mathcal{E}^{\mathrm{ap}}_{\mathrm{targ}}$. As mentioned in Section 3.4.3.3, this behaviour is typical of the algorithm.

In radio engineering terms, one iteration of the original Gerchberg-Saxton algorithm is described by (cf. (3.59))

$$\begin{aligned} G_i(u,v) &= \mathrm{FT}\{g_i(x,y)\} \\ G_i'(u,v) &= A_{\mathrm{m}}(u,v)e^{j\,\mathrm{phase}\{G_i(u,v)\}} \\ g_i'(x,y) &= \mathrm{IFT}\{G_i'(u,v)\} \\ g_{i+1}(x,y) &= |f_{\mathrm{d}}(x,y)|e^{j\,\mathrm{phase}\{g_i'(x,y)\}} \end{aligned} \tag{4.34}$$

Figures 4.12(b) and (c) depict error curves for the original Gerchberg-Saxton algorithm applied to two different models. The models are respectively defined by (4.32) and by

$$\begin{aligned} &\text{design 2;} \qquad \Omega_{\mathrm{pan}} = 0.019; \quad \psi_{\mathrm{pan}} = 1.0 \text{ rad.}; \\ &\psi_{\mathrm{quad}} = 1.0 \text{ rad.}; \quad \tau_{\mathrm{ran}} = 0.01; \\ &\Gamma_{\mathrm{ran}} = 0.001 = -60 \text{ dB}; \qquad \Gamma_{\mathrm{off}} = 0.002 \end{aligned} \tag{4.35}$$

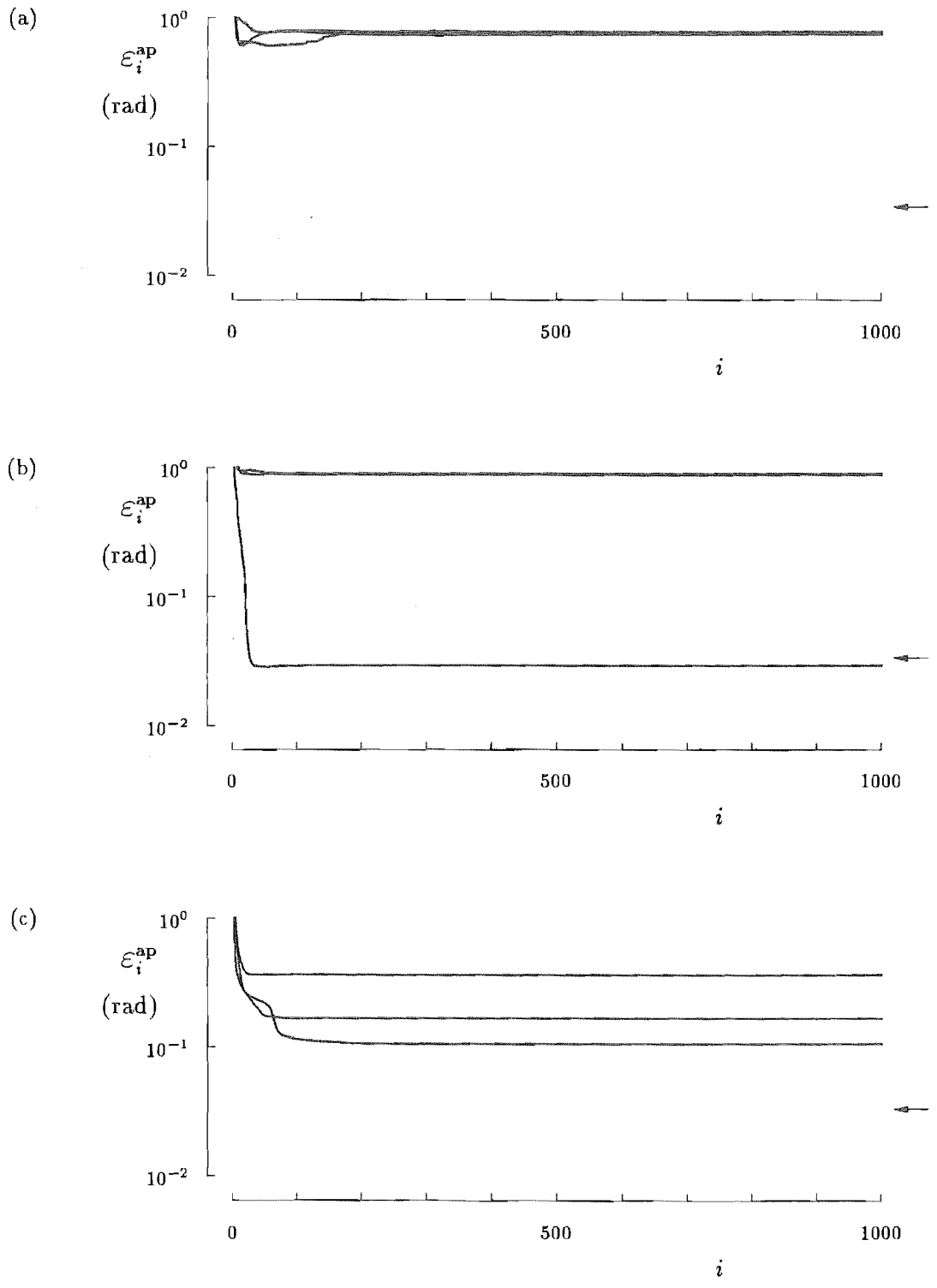


Figure 4.12 Error curves for algorithms of the error reduction type: (a) three runs of Fienup's error reduction algorithm for complex images applied to the model defined by (4.32); (b) three runs of the original Gerchberg-Saxton algorithm applied to the same model as in (a); (c) three runs of the original Gerchberg-Saxton algorithm applied to the model defined by (4.35). The arrows indicate the values of $\mathcal{E}_{\text{targ}}^{ap}$.

For each model, the original Gerchberg-Saxton algorithm was run three times for one thousand iterations using a different random starting phase distribution for each run.

Figure 4.12(b) shows that, for one of the three runs for the model defined by (4.32), the original Gerchberg-Saxton algorithm converged with $\mathcal{E}_{1000}^{\mathrm{ap}} < \mathcal{E}_{\mathrm{targ}}^{\mathrm{ap}}$. For the other model, however, Figure 4.12(c) indicates that all of the runs converged to values of $\mathcal{E}_{500}^{\mathrm{ap}}$ significantly greater than $\mathcal{E}_{\mathrm{targ}}^{\mathrm{ap}}$. It is my experience that the convergence characteristics displayed in Figure 4.12(c) typify the performance of the original Gerchberg-Saxton algorithm.

4.4.2 The CC algorithm

In this section, what is here called the CC algorithm is defined. The CC algorithm is based on what is here called the constant correction algorithm which is defined first.

The *constant correction algorithm* is a form of the variant of the Gerchberg-Saxton algorithm discussed in Section 3.4.3.2 [Gerchberg, 1986]. It is called the 'constant correction' algorithm because the normalized rms Fourier correction, defined by (3.68), remains constant from iteration to iteration. One iteration of the constant correction algorithm is described by (cf. (3.66)) [Milner *et al.*, 1987]

$$\begin{aligned} G_i(u,v) &= \mathrm{FT}\{g_i(x,y)\} \\ G_i'(u,v) &= A_{\mathrm{m}}(u,v)e^{j\,\mathrm{phase}\{G_i(u,v)\}}e^{j\left|\mathrm{phase}\{G_{i-1}'(u,v)\}-\mathrm{phase}\{G_i(u,v)\}\right|} \\ g_i'(x,y) &= \mathrm{IFT}\{G_i'(u,v)\} \\ g_{i+1}(x,y) &= |f_{\mathrm{d}}(x,y)|e^{j\,\mathrm{phase}\{g_i'(x,y)\}}e^{j\left|\mathrm{phase}\{g_i(x,y)\}-\mathrm{phase}\{g_i'(x,y)\}\right|} \end{aligned} \tag{4.36}$$

where $\mathrm{phase}\{G_{-1}'(u,v)\}$ is taken to be zero.

Figure 4.13(a) shows error curves for three runs, each of one thousand iterations, of the constant correction algorithm. The algorithm is applied to the model defined by (4.35) and invokes a different random starting phase distribution for each run. The graph indicates that even after one thousand iterations, the algorithm does not converge to a value of $\mathcal{E}_i^{\mathrm{ap}}$ close to that of $\mathcal{E}_{\mathrm{targ}}^{\mathrm{ap}}$.

The performance of the constant correction algorithm can be improved by applying it for several iterations, followed by several iterations of Fienup's error reduction algorithm. The *CC algorithm* is here defined to be 400 iterations of the constant correction algorithm, as defined by (4.36), followed by 100 iterations of Fienup's error reduction algorithm, as defined in (4.31).

The number of iterations in the CC algorithm is fixed in order to make the algorithm straightforward to implement. The decision to apply the constant correction and error reduction algorithms for 400 and 100 iterations, respectively, was made on the basis of many experiments involving these algorithms. Too few iterations of the constant correction algorithm result in the failure of many runs which would have converged satisfactorily after more iterations. Usually, application of more constant correction iterations does not result in a worse final value of $\mathcal{E}_i^{\mathrm{ap}}$. However, the larger the total number of iterations, the more computer time is required to implement the algorithm.

Intuitive reasoning, for why the CC algorithm is likely to be more successful than the constant correction algorithm, is now outlined. Consider the ideal situation in which $A_{\mathrm{m}}(u,v) = |F_{\mathrm{a}}(u,v)|$, implying that it is possible in principle for the modified Gerchberg-Saxton algorithm to converge to the exact solution, by which is meant that $f_{\mathrm{e}}(x,y)$, as generated by the algorithm, exactly equals the image-form of $f_{\mathrm{a}}(x,y)$. The aperture constraint invoked by the constant correction algorithm involves $|f_{\mathrm{e}}(x,y)|$.

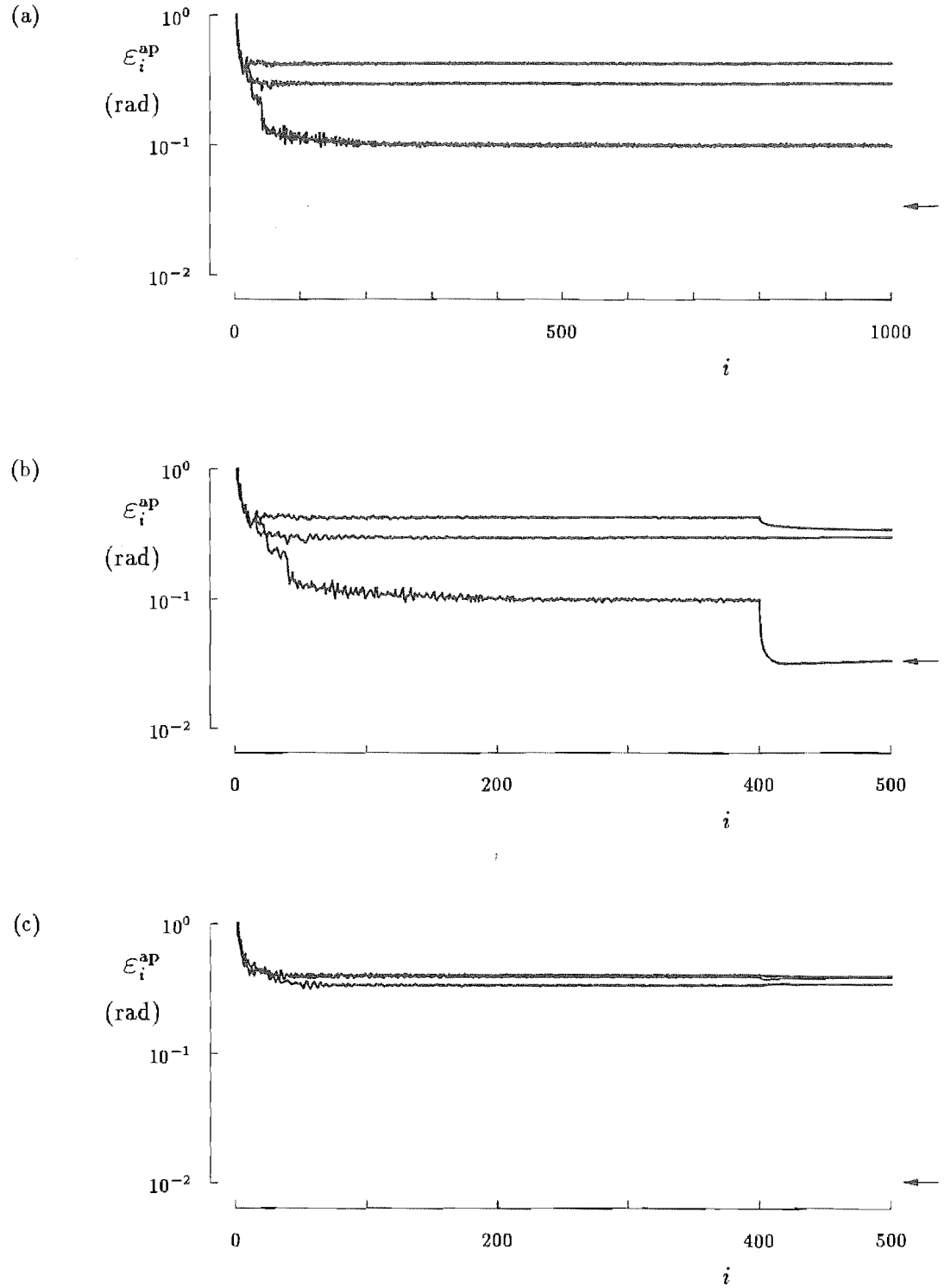


Figure 4.13 Error curves for the constant correction and CC algorithms: (a) three runs of the constant correction algorithm, each having 1000 iterations, applied to the model defined by (4.35); (b) three runs of the CC algorithm applied to the same model; (c) three runs of the CC algorithm applied to the model defined by (4.37). The arrows indicate the values of $\mathcal{E}_{targ}^{ap}$.

Because $|f_d(x,y)| \approx |f_a(x,y)|$, the constant correction algorithm can, at best, only generate a reasonably accurate solution, which can then be used as the starting aperture field distribution for Fienup's error reduction algorithm. The latter algorithm tends to converge rapidly when it is applied to a sufficiently accurate starting aperture field distribution. Because the aperture constraint for Fienup's error reduction algorithm involves only S^{aper}, this algorithm is free to generate an exact solution when presented with perfect data.

Figures 4.13(b) and (c) depict error curves for the CC algorithm applied to two models. The first model is defined by (4.35) and is the same as that to which the curves shown in Figure 4.13(a) refer. The second model is defined by

$$\begin{aligned} &\text{design 2;} \quad \psi_{\text{quad}} = 1.0 \text{ rad.;} \quad \tau_{\text{ran}} = 0.01; \\ &\Gamma_{\text{ran}} = 0.00032 = -70 \text{ dB;} \quad \Gamma_{\text{off}} = 0.00063 \end{aligned} \tag{4.37}$$

For each model the CC algorithm was applied for three different random starting phase distributions.

Comparison of Figures 4.13(a) and (b) shows that the CC algorithm represents an improvement over the constant correction algorithm: one of the three runs of the CC algorithm converges to a value of $\mathcal{E}_i^{\text{ap}}$ as low as $\mathcal{E}_{\text{targ}}^{\text{ap}}$, whereas none of the runs of the constant correction algorithm do so. One cannot always expect, however, that at least one of three runs of the CC algorithm converges to a value of $\mathcal{E}_{500}^{\text{ap}} \approx \mathcal{E}_{\text{targ}}^{\text{ap}}$. This is illustrated in Figure 4.13(c), which demonstrates that all three runs of the CC algorithm can sometimes stagnate with values of $\mathcal{E}_{500}^{\text{ap}}$ much greater than the value of $\mathcal{E}_{\text{targ}}^{\text{ap}}$. It is interesting to note that such poor convergence typically occurs when the measurement inaccuracies are relatively small (e.g. when $\Gamma_{\text{ran}} \leq -80$ dB).

It is my experience that the CC algorithm either converges to a value of $\mathcal{E}_{500}^{\text{ap}}$ very close to $\mathcal{E}_{\text{targ}}^{\text{ap}}$, or stagnates with a value of $\mathcal{E}_{500}^{\text{ap}}$ much greater than $\mathcal{E}_{\text{targ}}^{\text{ap}}$. When the measurement inaccuracies are not relatively small (e.g. when $\Gamma_{\text{ran}} \geq -70$ dB), it is likely that for at least one out of three runs of the CC algorithm $\mathcal{E}_{500}^{\text{ap}}$ will be as low as $\mathcal{E}_{\text{targ}}^{\text{ap}}$. For such measurement inaccuracies, the CC algorithm has the highest success rate of all the forms of the modified Gerchberg-Saxton algorithm which I have examined. Section 4.4.5 describes how the CC algorithm can be incorporated into a composite algorithm, which is as successful as the CC algorithm is when measurement inaccuracies are not relatively small, but is more successful than the CC algorithm when measurement inaccuracies are relatively small.

4.4.3 The HIO algorithm

In this section, what is here called the HIO algorithm is defined. It is based on a hybrid input-output type of algorithm, called the HIOGS algorithm, which is defined first.

Just as Fienup's hybrid input-output algorithm is a combination of his input-output and error reduction algorithms (Sec. 3.4.3.3), the HIOGS algorithm is a combination of Fienup's input-output algorithm and the original Gerchberg-Saxton algorithm. One iteration of the HIOGS algorithm is described by (cf. (3.73) and (4.34)) [Bates *et al.*, 1987]

$$\begin{aligned} G_i(u,v) &= \mathrm{FT}\{g_i(x,y)\} \\ G_i'(u,v) &= A_{\mathrm{m}}(u,v)e^{j\,\mathrm{phase}\{G_i(u,v)\}} \\ g_i'(x,y) &= \mathrm{IFT}\{G_i'(u,v)\} \\ g_{i+1}(x,y) &= \begin{cases} |f_{\mathrm{d}}(x,y)|e^{j\,\mathrm{phase}\{g_i'(x,y)\}} & \text{for } (x,y) \in S^{\mathrm{aper}} \\ g_i(x,y) - 0.5g_i'(x,y) & \text{elsewhere} \end{cases} \end{aligned} \tag{4.38}$$

In words, the HIOGS algorithm is the original Gerchberg-Saxton algorithm, but with Fienup's input-output constraint applied to the part of the aperture field distribution which lies outside S^{aper}. For reasons given in Section 3.4.3.3, $g_i'(x,y)$, instead of $g_i(x,y)$, is the estimate of the image-form of $f_{\mathrm{a}}(x,y)$. However, it is nevertheless true that $\mathcal{E}_i^{\mathrm{ap}}$, which indicates the phase error of $g_i(x,y)$, remains a valid error measure because $\mathrm{phase}\{g_i(x,y)\} = \mathrm{phase}\{g_i'(x,y)\}$ in the region over which $\mathcal{E}^{\mathrm{ap}}$ is calculated. Comparison of (3.73) with (4.38), indicates that the feedback parameter β has been set to 0.5.

Error curves are presented in Figure 4.14(a) for the HIOGS algorithm applied to data generated from the model defined in (4.37). The algorithm was run three times using different random starting phase distributions, and for each run, was applied for one thousand iterations. Notice how, after $\mathcal{E}^{\mathrm{ap}}$ drops to a minimum value, it tends to increase with more iterations.

Following the same reasoning as given in Section 4.4.2 for the CC algorithm, what is here called the *HIO algorithm* is defined to consist of 400 iterations of the HIOGS algorithm followed by 100 iterations of Fienup's error reduction algorithm. The latter is defined in (4.31).

Figures 4.14(b) and (c) show the results of applying the HIO algorithm to the models defined by (4.37) and (4.35) respectively. For each model, the algorithm is applied three times, utilizing a different random starting phase distribution each time. Note that, at the completion of the algorithm, the value of $\mathcal{E}_{500}^{\mathrm{ap}}$ is typically approximately equal to the minimum value of $\mathcal{E}_i^{\mathrm{ap}}$ taken over all of the iterations.

As indicated in the figures, the smallest value of $\mathcal{E}^{\mathrm{ap}}$ resulting from three runs of the algorithm is typically no more than twice the target value of $\mathcal{E}^{\mathrm{ap}}$, especially when $\Gamma_{\mathrm{ran}} \geq -70$ dB. In my experience, it is very rare for all three runs of the HIO algorithm to stagnate with a value of $\mathcal{E}_{500}^{\mathrm{ap}}$ more than about twice the value of $\mathcal{E}_{\mathrm{targ}}^{\mathrm{ap}}$. On occasion, especially when the measurement inaccuracies are relatively small, the HIO algorithm converges to a value of $\mathcal{E}_{500}^{\mathrm{ap}}$ which is very close to $\mathcal{E}_{\mathrm{targ}}^{\mathrm{ap}}$. This fairly consistent convergence behaviour of the HIO algorithm contrasts with many other forms of the modified Gerchberg-Saxton algorithm, such as the CC algorithm, which often tend to converge well when applied to some kinds of data, but converge badly when applied to other kinds. Section 4.4.5 describes a composite algorithm which incorporates the HIO algorithm. The degree of convergence of the composite algorithm is almost never worse than that of the HIO algorithm.

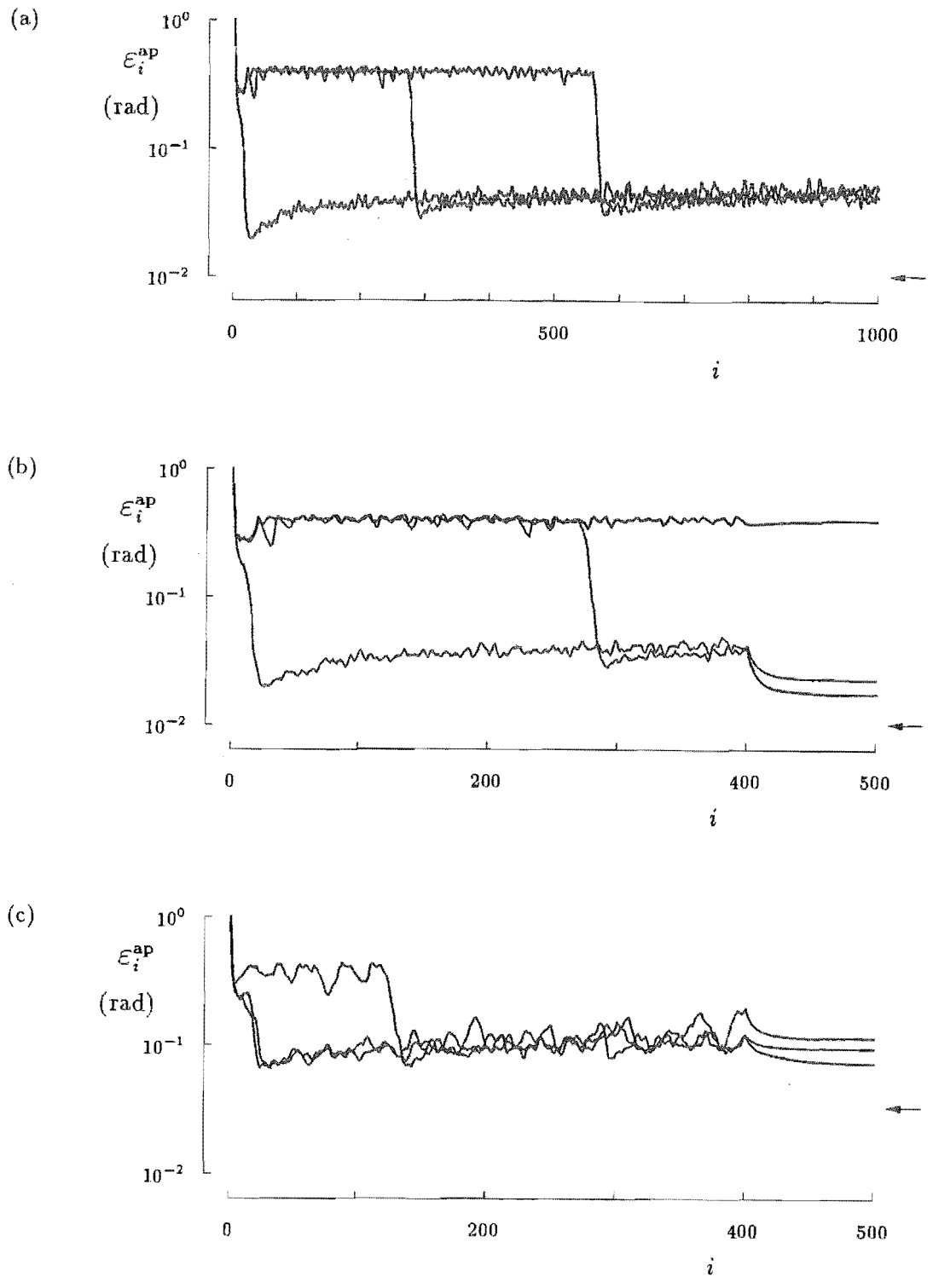


Figure 4.14 Error curves for the HIOGS and HIO algorithms: (a) three runs of the HIOGS algorithm, each having 1000 iterations, applied to the model defined by (4.37); (b) three runs of the HIO algorithm applied to the same model; (c) three runs of the HIO algorithm applied to the model defined by (4.35). The arrows indicate the values of $\mathcal{E}_{\text{targ}}^{ap}$.

4.4.4 Choice of starting aperture distribution

In all of the examples of the modified Gerchberg-Saxton algorithm presented so far in this chapter, the form of the starting aperture field distribution $g_1(x, y)$ is that defined by (4.33). In this section, alternative forms for $g_1(x, y)$ are discussed.

The more closely $g_1(x, y)$ resembles $f_a(x, y)$, the faster the algorithm converges. This is illustrated with the aid of Figure 4.15, which shows error curves for the CC and HIO algorithms applied to the model defined by (4.35) with $g_1(x, y)$ having the form

$$g_1(x, y) = |f_d(x, y)| e^{j[\text{phase}\{f_a(x,y)\} + \psi_{\text{start}} \text{ ran}(x,y)]} \tag{4.39}$$

where ψ_{start} is a parameter characterizing the difference in phase between $g_1(x, y)$ and $f_a(x, y)$. Note that the error curves for the CC and HIO algorithms can usually be distinguished by the way that they fluctuate (e.g. compare Figs. 4.13 and 4.14).

It is instructive to consider the case where $\psi_{\text{start}} = 0$ rad., which is depicted in Figure 4.15(a). Although $\mathcal{E}_1^{\text{ap}}$ must necessarily be zero, $\mathcal{E}_{500}^{\text{ap}}$ for both the CC and the HIO algorithms is non-zero. This is because the measurement inaccuracy inherent in $A_m(u, v)$ ensures that there is no exact solution to the phase problem (see Sec. 3.4.2.4).

Figures 4.15(b) and (c) show the error curves corresponding to values of ψ_{start} equal to 1.0 and $\pi/\sqrt{3}$ radians respectively. For both sets of curves the CC and HIO algorithms were each run three times, using a different set of random numbers when calculating $g_1(x, y)$ for each run. Note that (4.39) defines $g_1(x, y)$. Note also that, when $\psi_{\text{start}} = \pi/\sqrt{3}$, (4.39) reduces to (4.33). Comparison of Figure 4.15(b) with Figure 4.15(c) shows that, the more accurately $g_1(x, y)$ approximates $f_a(x, y)$, the faster does the algorithm converge over the first few iterations, for the HIO algorithm, and over the first few tens of iterations, for the CC algorithm. However, the final values of $\mathcal{E}_i^{\text{ap}}$ are not significantly different for the two algorithms.

If a sufficiently accurate estimate of phase$\{f_a(x, y)\}$ is available, the required number of modified Gerchberg-Saxton iterations can be reduced because of its faster rate of convergence. In my early work [Gardenier *et al.*, 1986b; Gardenier *et al.*, 1986c] I found that the original Gerchberg-Saxton algorithm converges well only when the rms phase difference between $g_1(x, y)$ and $f_a(x, y)$ is less than about 0.7 rad. In order to find a suitable $g_1(x, y)$, I ran the algorithm several times, each time with phase$\{g_1(x, y)\}$ set to a different radially quadratic distribution, until the algorithm manifested convergence instead of stagnation. This trial and error approach, to finding a $g_1(x, y)$ which is close enough to $f_a(x, y)$, has the disadvantage that the algorithm must be run many times before a suitable $g_1(x, y)$ can be found.

Another way of obtaining an estimate of $f_a(x, y)$ is inherent in the algorithm described by Anderson *et al.* [1988]. They have available a noisy measured copolar phase pattern as well as a more accurately measured copolar amplitude pattern. Their algorithm is similar to the modified Gerchberg-Saxton when $g_i(x, y)$ is the inverse Fourier transform of the measured (amplitude and phase) copolar pattern. When applied to such data, of the modified Gerchberg-Saxton algorithm can be thought of as a means for improving the measured estimate of the copolar phase pattern [Anderson *et al.*, 1988].

This thesis is, however, mainly concerned with the problem of estimating the image-form of $f_a(x, y)$ when no measurement of the copolar phase pattern is available. The best available estimate of $f_a(x, y)$ is then $f_d(x, y)$. Figure 4.16 depicts the error curves for the CC and HIO algorithms applied to data generated from the model defined by (4.35) when $g_1(x, y) = f_d(x, y)$. For this particular model, the CC algorithm converges faster than when $g_1(x, y)$ has the form described in (4.33). However, my experience

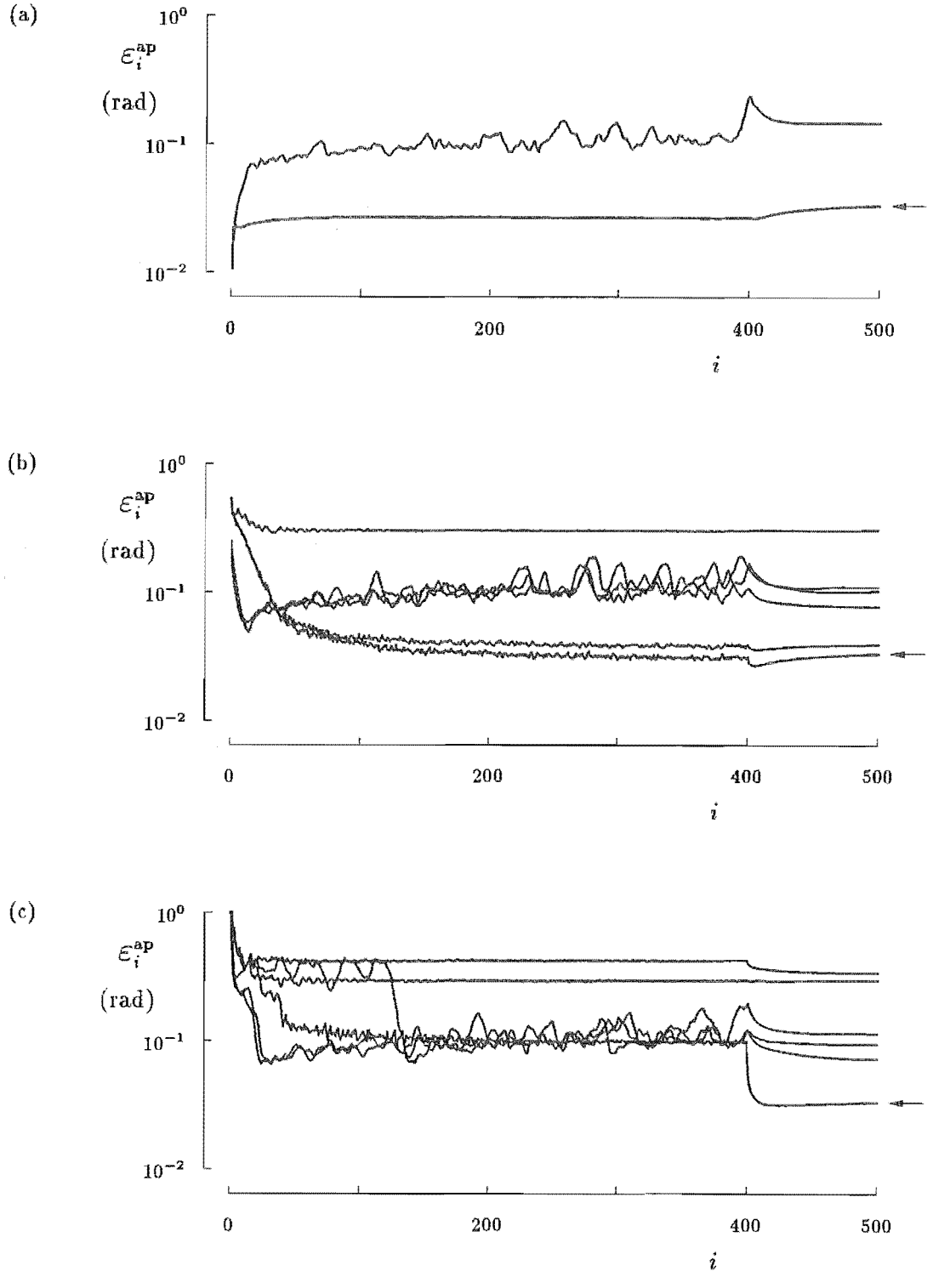


Figure 4.15 Error curves for the CC and HIO algorithms applied with a starting aperture distribution $g_1(x, y)$ defined by (4.39): (a) one run of each algorithm with $\psi_{start} = 0$ rad.; (b) three runs of each algorithm with $\psi_{start} = 1$ rad.; (c) three runs of each algorithm with $\psi_{start} = \pi/\sqrt{3}$ rad. The arrows indicate the values of $\mathcal{E}_{targ}^{ap}$. The model to which the algorithms are applied is defined by (4.35).

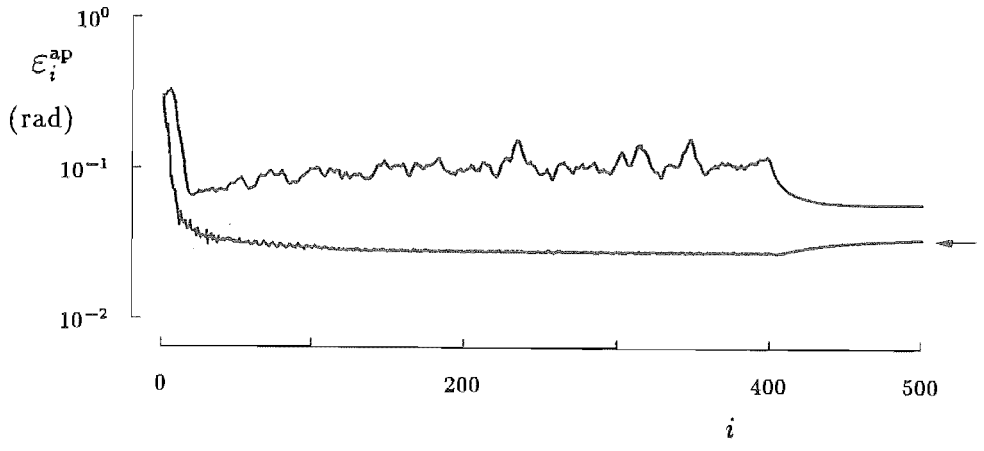


Figure 4.16 Error curves for the CC and HIO algorithms applied with a starting aperture distribution of $g_1(x,y) = f_{\mathrm{d}}(x,y)$. The arrows indicate the values of $\mathcal{E}^{\mathrm{ap}}_{\mathrm{targ}}$. The model to which the algorithms were applied is defined by (4.35).

of applying the modified Gerchberg-Saxton algorithm to many models suggests that there is often little difference between the convergence properties of the algorithm when $g_1(x,y) = f_{\mathrm{d}}(x,y)$ and when $g_1(x,y)$ has a random starting phase distribution as expressed by (4.33).

The advantage of using a random starting phase distribution, as opposed to a particular estimate of phase$\{f_{\mathrm{a}}(x,y)\}$, is that the algorithm can be run several times with independent starting phase distributions. In this way, if the initial run fails to converge sufficiently, the algorithm can be run more times in an attempt to find a starting phase distribution which causes the algorithm to converge adequately for whatever application is envisaged.

4.4.5 The composite algorithm

My experience is that, when applied to an arbitrary model, either the CC algorithm or the HIO algorithm usually converges as well as, or better than, any other form of the modified Gerchberg-Saxton algorithm that I have experimented with. However, it often takes up to three runs of either algorithm, using a different random starting phase distribution for each run, to produce a run which converges satisfactorily.

With the above discussion in mind, it is convenient to define what is here called the *composite algorithm*, which is the form of the modified Gerchberg-Saxton algorithm applied in all examples presented in the remainder of thesis. It is defined by the following steps:

1. Run the CC algorithm three times using a starting aperture field distribution of the form described by (4.33) with a different random starting phase distribution for each run.

2. Run the HIO algorithm three times using the same starting aperture field distributions as were used for the CC algorithm.

3. Out of the total of six runs, choose the one for which the value of $\mathcal{E}^{\mathrm{fa}}_{500}$ is smallest. The estimate $f_{\mathrm{e}}(x,y)$ generated by the composite algorithm is then taken to be $g_{500}(x,y)$ for the chosen run.

Note that, in step (3), $\mathcal{E}^{\mathrm{fa}}_{500}$ is invoked as the measure of convergence instead of $\mathcal{E}^{\mathrm{ap}}_{500}$, which is used in the previous sections. This is because it is envisaged that the composite algorithm is a form of the modified Gerchberg-Saxton algorithm that could be applied in practice. The error measure $\mathcal{E}^{\mathrm{fa}}_{i}$ could be computed in practice but $\mathcal{E}^{\mathrm{ap}}_{i}$ could not (Sec. 4.3). The comparative significance of these two errors is discussed in Section 4.6.

In the discussion in Section 3.4.3.3 of Fienup's algorithms, it is mentioned that the relationship between the quality of the image generated by an algorithm and the corresponding value of the image error is different for the different algorithms. This problem does not occur when comparing the images generated by the CC and HIO algorithms on the basis of $\mathcal{E}^{\mathrm{fa}}_{500}$, because both of these algorithms finish with 100 iterations of Fienup's error reduction algorithm.

The reason for running both the CC and the HIO algorithms in the composite algorithm is to gain the advantages of each. Except when the measurement inaccuracies are very small, one of the CC algorithm runs usually converges better than all of the HIO algorithm runs. An example of this is illustrated by comparing Figure 4.13(b) with Figure 4.14(c) which display error curves of, respectively, the CC and HIO algorithms applied to one particular model. However, occasionally all of the CC algorithm runs stagnate with a value of $\mathcal{E}^{\mathrm{ap}}$ much greater than $\mathcal{E}^{\mathrm{ap}}_{\mathrm{targ}}$. On the other hand, as pointed out in Section 4.4.3, the best of the runs of the HIO algorithm tends to consistently converge to a value of $\mathcal{E}^{\mathrm{ap}}_{500}$ which is about twice that of $\mathcal{E}^{\mathrm{ap}}_{\mathrm{targ}}$. Therefore, when the CC algorithm runs fail to converge sufficiently, the HIO algorithm runs serve as a back up. Figures 4.13(c) and 4.14(b), which show error curves of the CC and HIO algorithms applied to one particular model, constitute an example of the convergence of the HIO algorithm being better than that of the CC algorithm.

4.4.6 Alternative forms of the modified Gerchberg-Saxton algorithm

The following five sections introduce five alternative versions, besides those already discussed in Sections 4.4.1 to 4.4.5, of the modified Gerchberg-Saxton algorithm. My experience with these versions suggest that they are not as useful as the ones described in Sections 4.4.1 to 4.4.5. However, they are included here for completeness and as examples of different ways of applying the constraints in the modified Gerchberg-Saxton algorithm.

4.4.6.1 Local well avoidance

One algorithm was developed in an attempt to help the original Gerchberg-Saxton algorithm out of the local 'well' (see Sec. 3.4.3.2) after it has stagnated. One iteration of the original Gerchberg-Saxton algorithm is defined by (4.34). However, it is here convenient to rewrite the last equation of (4.34) as

$$g_{i+1}(x,y) = g'_i(x,y) + \Delta g_i(x,y) \tag{4.40}$$

where $\Delta g_i(x,y)$ is the change that occurs in the aperture plane and is defined by

$$\Delta g_i(x,y) = \big||f_{\mathrm{d}}(x,y)| - |g'_i(x,y)|\big|\, e^{j\,\mathrm{phase}\{g'_i(x,y)\}} \tag{4.41}$$

Note that, in the same way that (3.61) to (3.64) show that $\mathcal{E}_i^{\mathrm{F}}$ is a measure of the degree of convergence of the original Gerchberg-Saxton algorithm, it can equivalently be shown that the rms value of $|\Delta g_i(x,y)|$ is also a measure of the algorithm's degree of convergence. In Section 3.4.3.1 it is demonstrated that stagnation occurs when $g_{i+1}(x,y)$ is almost equal to $g_i(x,y)$. Therefore, a way to get out of stagnation is to alter (4.40) so that $g_{i+1}(x,y)$ is different from $g_i(x,y)$. Since its amplitude is kept equal to that of $|f_{\mathrm{d}}(x,y)|$, it is only with regard to its phase that $g_{i+1}(x,y)$ can differ from $g_i(x,y)$. It seems reasonable to make the phase difference between $g_{i+1}(x,y)$ and $g_i(x,y)$ approximately proportional to $|\Delta g_i(x,y)|$. Therefore, only small changes need to be made to the original Gerchberg-Saxton algorithm. An appropriate modification to the original Gerchberg-Saxton algorithm is to replace (4.40) with (cf. (4.34))

$$g_{i+1}(x,y) = |f_{\mathrm{d}}(x,y)| e^{j\,\mathrm{phase}\{g_i'(x,y)+j\Delta g_i(x,y)\}} \tag{4.42}$$

It is suggested that (4.42) should replace (4.40) only when the original Gerchberg-Saxton algorithm stagnates. This would mean that the replacement would apply for only a small number of iterations: until the algorithm is out of the local 'well'. Thereafter, (4.40) would be invoked so that the original Gerchberg-Saxton algorithm can continue. However, my experience with this approach is that, after invoking (4.42), the original Gerchberg-Saxton algorithm rapidly stagnates once again, seemingly no closer to a solution. The lesson learned from this is that the original Gerchberg-Saxton algorithm should only be invoked when the starting distribution $g_1(x,y)$ is already close to the solution. Then, assuming that the local well is in fact the global well, the original Gerchberg-Saxton algorithm is guaranteed to generate a more accurate estimate of the solution.

4.4.6.2 The phase relaxation algorithm

Another approach to modifying the Gerchberg-Saxton algorithm is to introduce a relaxation parameter (the constant μ introduced below) into the aperture phase distributions generated at each iteration of the original Gerchberg-Saxton algorithm. Instead of setting $\mathrm{phase}\{g_{i+1}(x,y)\}$ equal to the right side of the final equation of (4.34) (which defines the original Gerchberg-Saxton algorithm), it is instead set equal to a linear combination of the right side of the final equation of (4.34) and of $g_i(x,y)$. One iteration, of what is here called the *phase relaxation algorithm*, is defined by (4.34) with the last equation replaced by

$$g_{i+1}(x,y) = |f_{\mathrm{d}}(x,y)| e^{j[\mu\,\mathrm{phase}\{g_i'(x,y)\}+(1-\mu)\,\mathrm{phase}\{g_i(x,y)\}]} \tag{4.43}$$

where μ is the relaxation constant, which lies between 0 and 1. When $\mu = 1$ this algorithm reduces to the original Gerchberg-Saxton algorithm. Dr. W. Richard Fright (when he was a post-doctoral fellow at the University of Canterbury) found this approach helpful for his work in acoustic microscopy [described by Fright *et al.*, 1989] which involved a one-dimensional Fourier phase problem.

4.4.6.3 Constraints involving thresholds

When the data $A_{\mathrm{m}}(u,v)$ and $|f_{\mathrm{d}}(x,y)|$ are inaccurate, the original Gerchberg-Saxton algorithm can never exactly converge to a solution $f_{\mathrm{e}}(x,y)$ for which $|f_{\mathrm{e}}(x,y)| = |f_{\mathrm{d}}(x,y)|$ and $|F_{\mathrm{e}}(u,v)| = A_{\mathrm{m}}(u,v)$ (Sec. 3.4.2.4). This implies that the aperture constraint and the far field constraint are incompatible. One way of relaxing the constraints, so that

they become compatible is to impose the conditions $|f_e(x,y)| = |f_d(x,y)| \pm \Gamma^a_{thres}$ and $|F_e(u,v)| = A_m(u,v) \pm \Gamma^f_{thres}$, where the real numbers Γ^a_{thres} and Γ^f_{thres} are thresholds. Ideally, the thresholds are chosen so that $f_e(x,y)$ would meet the relaxed constraints where it is identical to $f_a(x,y)$. However, if the thresholds are too large there may be many different images $f_e(x,y)$ which meet the constraints. One iteration of what is here called the *threshold algorithm* is defined by

$$\begin{aligned}
G_i(u,v) &= \mathrm{FT}\{g_i(x,y)\} \\
G'_i(u,v) &= \begin{cases} A_m(u,v)e^{j\,\mathrm{phase}\{G_i(u,v)\}} & \text{if } ||G_i(u,v)| - A_m(u,v)| > \Gamma^f_{thres} \\ G_i(u,v) & \text{otherwise} \end{cases} \\
g'_i(x,y) &= \mathrm{IFT}\{G'_i(u,v)\} \\
g_{i+1}(x,y) &= \begin{cases} |f_d(x,y)|e^{j\,\mathrm{phase}\{g'_i(x,y)\}} & \text{if } ||g'_i(x,y)| - |f_d(x,y)|| > \Gamma^a_{thres} \\ g'_i(x,y) & \text{otherwise} \end{cases}
\end{aligned} \tag{4.44}$$

Another algorithm in which thresholds are applied was suggested by Gabor T. Herman (when he visited the University of Canterbury from the Medical Imaging Group, Department of Radiology, Hospital of the University of Pennsylvania). It is here called the *reflection algorithm* and one of its iterations is described by (cf. (4.44))

$$\begin{aligned}
G_i(u,v) &= \mathrm{FT}\{g_i(x,y)\} \\
G'_i(u,v) &= \begin{cases} A_m(u,v)e^{j\,\mathrm{phase}\{G_i(u,v)\}} \\ \qquad \text{if } ||G_i(u,v)| - A_m(u,v)| > 2\,\Gamma^f_{thres} \\ [2A_m(u,v) + 2\,\Gamma^f_{thres} - |G_i(u,v)|]e^{j\,\mathrm{phase}\{G_i(u,v)\}} \\ \qquad \text{if } \Gamma^f_{thres} < [|G_i(u,v)| - A_m(u,v)] < 2\,\Gamma^f_{thres} \\ [2A_m(u,v) - 2\,\Gamma^f_{thres} - |G_i(u,v)|]e^{j\,\mathrm{phase}\{G_i(u,v)\}} \\ \qquad \text{if } -2\,\Gamma^f_{thres} < [|G_i(u,v)| - A_m(u,v)] < -\Gamma^f_{thres} \\ G_i(u,v) \\ \qquad \text{otherwise} \end{cases} \\
g'_i(x,y) &= \mathrm{IFT}\{G'_i(u,v)\} \\
g_{i+1}(x,y) &= \begin{cases} |f_d(x,y)|e^{j\,\mathrm{phase}\{g'_i(x,y)\}} \\ \qquad \text{if } ||g'_i(x,y)| - |f_d(x,y)|| > 2\,\Gamma^a_{thres} \\ [2|f_d(x,y)| + 2\,\Gamma^a_{thres} - |g'_i(x,y)|]e^{j\,\mathrm{phase}\{g'_i(x,y)\}} \\ \qquad \text{if } \Gamma^a_{thres} < [|g'_i(x,y)| - |f_d(x,y)|] < 2\,\Gamma^a_{thres} \\ [2|f_d(x,y)| - 2\,\Gamma^a_{thres} - |g'_i(x,y)|]e^{j\,\mathrm{phase}\{g'_i(x,y)\}} \\ \qquad \text{if } -2\,\Gamma^a_{thres} < [|g'_i(x,y)| - |f_d(x,y)|] < -\Gamma^a_{thres} \\ g'_i(x,y) \\ \qquad \text{otherwise} \end{cases}
\end{aligned} \tag{4.45}$$

The second equation of (4.45) implies that if $|G_i(u,v)|$ exceeds $(A_m(u,v) + \Gamma^f_{thres})$ by, say, d, and provided d is less than Γ^f_{thres}, then $|G'_i(u,v)|$ is chosen such that it is less than $(A_m(u,v) + \Gamma^f_{thres})$ by d. The reflection algorithm therefore has the advantage that if $G_i(u,v)$ almost meets the far field constraint then $G'_i(u,v)$ is close to $G_i(u,v)$.

Although I have not found the reflection algorithm to be as successful as the composite algorithm (Sec. 4.4.5), I have usefully incorporated aspects of it into the extrapolating composite algorithm (Sec. 4.7.3.3) which is an extension of the composite algorithm suitable for application to truncated far field data (Sec. 4.7.3).

4.4.6.4 Another input-output algorithm

Section 3.4.3.3 describes Fienup's input-output algorithm which can be utilized either for positive images or for complex images whose support is known. Fienup [1980] has also developed an input-output algorithm for complex images whose amplitude distribution is known. Recall from Section 3.4.3.3 that the input output algorithm regards the first three equations of (4.31) as describing a nonlinear process, with an input $g_i(x,y)$ and an output $g_i'(x,y)$. For a complex image whose amplitude is known, the purpose of the input-output algorithm is to find a $g_i(x,y)$ which drives the amplitude of $g_i'(x,y)$ to equal the known amplitude distribution. It turns out [Fienup, 1980] that there are many ways of implementing such an algorithm. One iteration of the simplest implementation is described by (4.31) with its final equation replaced by

$$g_{i+1}(x,y) = g_i(x,y) + \beta \Delta g_i(x,y) \tag{4.46}$$

where $\Delta g_i(x,y)$ is defined in (4.41). Note that (4.46) reduces to (3.72) in regions of the aperture plane outside S^{f_a}.

It is my experience that the input-output algorithm described in this section and each of the versions of the modified Gerchberg-Saxton algorithm described in Sections 4.4.6.1 through 4.4.6.3 converge as well as, or even better than, the composite algorithm, when applied to data generated from specific models. However, my experience with applying these algorithms to many different models is that, in general, the constant correction algorithm tends to converge to more a accurate estimate of the image-form of $f_a(x,y)$ than does any of the other algorithms with which I have worked.

4.5 A WORKED EXAMPLE

The results presented in Sections 4.7 to 4.9 describe the performance of the composite algorithm (Sec. 4.4.5) in terms of error measures (Sec. 4.3). However, to provide a better understanding of the composite algorithm (Sec. 4.4.5), this section works through, in detail, an example of the algorithm applied to the data generated from a particular model.

The particular model invoked for this worked example is here called the *basic model.* It is described by (4.35), which is repeated here for emphasis:

$$\begin{aligned} &\text{design 2;} \qquad \Omega_{\text{pan}} = 0.019; \quad \psi_{\text{pan}} = 1.0 \text{ rad.;} \\ &\psi_{\text{quad}} = 1.0 \text{ rad.;} \quad \tau_{\text{ran}} = 0.01; \\ &\Gamma_{\text{ran}} = 0.001 = -60 \text{ dB;} \qquad \Gamma_{\text{off}} = 0.002 \end{aligned} \tag{4.47}$$

The design copolar aperture field distribution $f_{\text{d}}(x, y)$ is displayed in Figure 4.17. As for all fields depicted in this section, diagonal cuts through the copolar aperture field amplitude and phase distributions are shown. Position along the diagonal is indicated by ξ, which is defined in (4.8). Cuts, along the u axis, through the corresponding copolar far field amplitude pattern and through the design envelope $\Lambda_{\text{d}}(u, v)$ are also depicted. The amplitude pattern and $\Lambda_{\text{d}}(u, v)$ are always plotted in decibels relative to their respective peak values.

The actual copolar aperture field distribution $f_{\text{a}}(x, y)$ is shown in Figure 4.18. Note that phase$\{f_{\text{a}}(x, y)\}$ has large random-like values in the region blocked by the subreflector. This is because the field in this region is entirely due to the scattered field characterized by τ_{ran} (see (4.14)). For comparison with later figures, the centre 33 by 33 samples of phase$\{f_{\text{a}}(x, y)\}$ are depicted in Figure 4.19, in which the effect of the radially quadratic term (characterized by ψ_{quad}) and the panel term (characterized by ψ_{pan} and Ω_{pan}) are clearly seen. Comparison of $|F_{\text{a}}(u, v)|$ (Fig. 4.18(c)) with $|F_{\text{d}}(u, v)|$ (Fig. 4.17(c)) shows the effect on the amplitude pattern of the aperture field deviations. The measured copolar far field amplitude pattern $A_{\text{m}}(u, v)$ and the actual copolar far field amplitude pattern $|F_{\text{a}}(u, v)|$ are depicted in Figure 4.18(c). The effect of the -60 dB far field measurement noise can be seen by comparing the graphs of $|F_{\text{a}}(u, v)|$ with $A_{\text{m}}(u, v)$ in Figure 4.18(c). The level of $A_{\text{m}}(u, v)$ exceeds the design envelope $\Lambda_{\text{d}}(u, v)$ because of increased sidelobe levels and a wider beamwidth. The value of E_{m} for this model is 3.21 dB.

The input data to the composite algorithm are $|f_{\text{d}}(x, y)|$ and $A_{\text{m}}(x, y)$. Following steps (1) and (2) of the composite algorithm (Sec. 4.4.5), the CC and HIO algorithms were each applied three times to these data, using a different random starting phase distribution each time. The values of $\mathcal{E}_{500}^{\text{fa}}$ generated by the six runs were 1.69×10^{-3}, 7.84×10^{-4}, 1.75×10^{-3}, 9.25×10^{-4}, 8.26×10^{-4} and 9.08×10^{-4}. The smallest of these values corresponds to one of the CC algorithm runs. This run is therefore chosen in step (3) of the composite algorithm and is now analysed in more detail.

The error curves, of both $\mathcal{E}_i^{\text{fa}}$ and $\mathcal{E}_i^{\text{ap}}$, pertaining to this chosen run are depicted in Figure 4.20. It can be seen that the error reduction part of the algorithm started to stagnate after $\mathcal{E}_i^{\text{fa}}$ fell below the measurement noise level of $\Gamma_{\text{ran}} = 0.001$. This is unavoidable, since the algorithm cannot be expected to generate an estimate $|F_{\text{e}}(u, v)|$ of the copolar far field amplitude pattern which is more accurate than $A_{\text{m}}(u, v)$.

Recall from Section 4.4.2 that $g_i(x, y)$ is the i^{th} estimate of the image-form of $f_{\text{a}}(x, y)$ generated by the CC algorithm. Figure 4.21 presents plots of the centre 33 by 33

samples of phase$\{g_i(x,y)\}$ for $i = 1, 50$ and 400. This includes the random starting phase distribution phase$\{g_1(x,y)\}$ and the phase distribution at the final constant correction algorithm iteration (see Sec. 4.4.2). Recall that the final aperture distribution $g_{500}(x,y)$ generated by this run of the CC algorithm is taken to be the estimate $f_e(x,y)$, of the image-form of the actual copolar aperture field distribution $f_a(x,y)$, generated by the composite algorithm. Its phase is plotted in Figure 4.22(a). By comparing this figure with Figure 4.19, it can be seen that phase$\{f_e(x,y)\}$ approximates an upside-down, rotated version of phase$\{f_a(x,y)\}$. This implies that $\tilde{f}_e(x,y)$, instead of $f_e(x,y)$, is an approximation of $f_a(x,y)$. To make this obvious, phase$\{\tilde{f}_e(x,y)e^{j0.619}\}$ is plotted in Figure 4.22(b). It is seen to be very similar to phase$\{f_a(x,y)\}$. In particular, the quadratic phase term due to defocusing and the effect of the displaced panel can be readily identified.

Cuts through $f_e(x,y)$ and $|F_e(u,v)|$ are shown in Figure 4.23. This figure indicates the constraints that were applied in the final iterations of the CC algorithm: $f_e(x,y)$ is a copolar aperture field distribution, which is zero outside the aperture support S^{aper}, and whose copolar far field amplitude pattern is as similar as possible, in the sense of minimizing $\mathcal{E}_i^{fa}$, to the measured data $A_m(u,v)$, which is depicted in Figure 4.18(c).

The correction process, based on information in phase$\{f_e(x,y)\}$, is modelled by (4.25). The corrected copolar aperture field distribution $f_c(x,y)$ is plotted in Figure 4.24. Note that its amplitude is equal to that of $|f_a(x,y)|$ while its phase is the difference in phase between $f_a(x,y)$ and $\tilde{f}_e(x,y)$ less a constant. The plot of phase$\{f_c(x,y)\}$, shown in Figure 4.22(c), provides an indication of how accurately phase$\{\tilde{f}_e(x,y)\}$ approximates phase$\{f_a(x,y)\}$. Note that phase$\{f_c(x,y)\}$ has a random appearance, which is, in my experience, typical for the composite algorithm applied to an arbitrary model. An encouraging aspect of this worked example is that the corrected copolar far field amplitude pattern $|F_c(u,v)|$ nowhere exceeds $\Lambda_d(u,v)$, as can be seen from Figure 4.24(c). Therefore, the corresponding value of E_c is 0 dB, when the composite algorithm is applied to data generated from the basic model.

For the examples presented throughout the remainder of this chapter, the convergence of the composite algorithm, when applied to data generated from a particular model, is indicated by $\mathcal{E}^{ap}$, E_m and E_c. For comparison with these other examples, the values, pertaining to this worked example, of $\mathcal{E}^{ap}$, E_m and E_c are plotted in Figure 4.25.

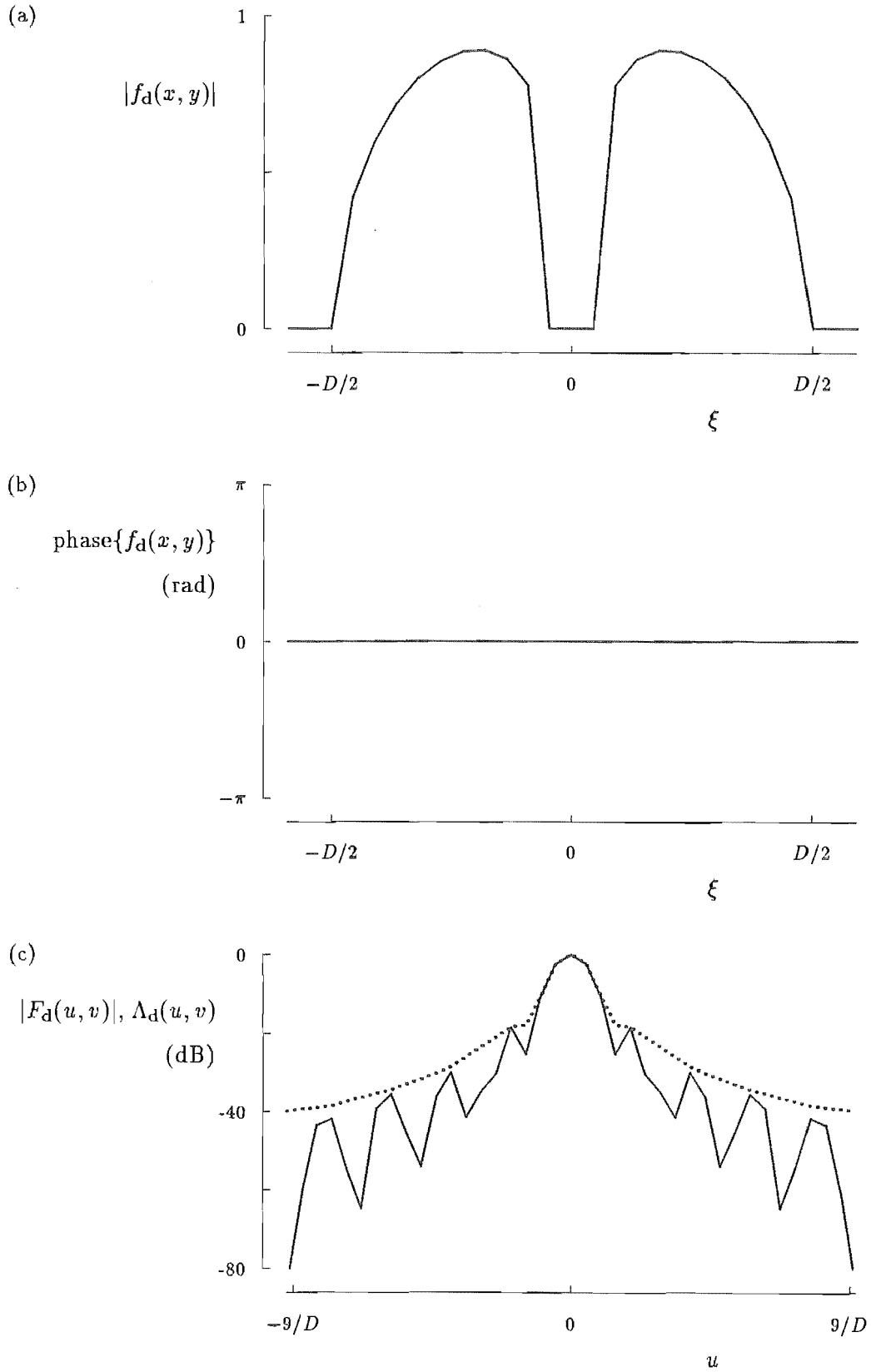


Figure 4.17 Design copolar fields for the worked example: (a) design copolar aperture field amplitude distribution; (b) design copolar aperture field phase distribution; (c) design copolar far field amplitude pattern (solid curve) and design envelope (dotted curve). The variable ξ, which identifies position along a diagonal in the aperture plane, is defined in (4.8).

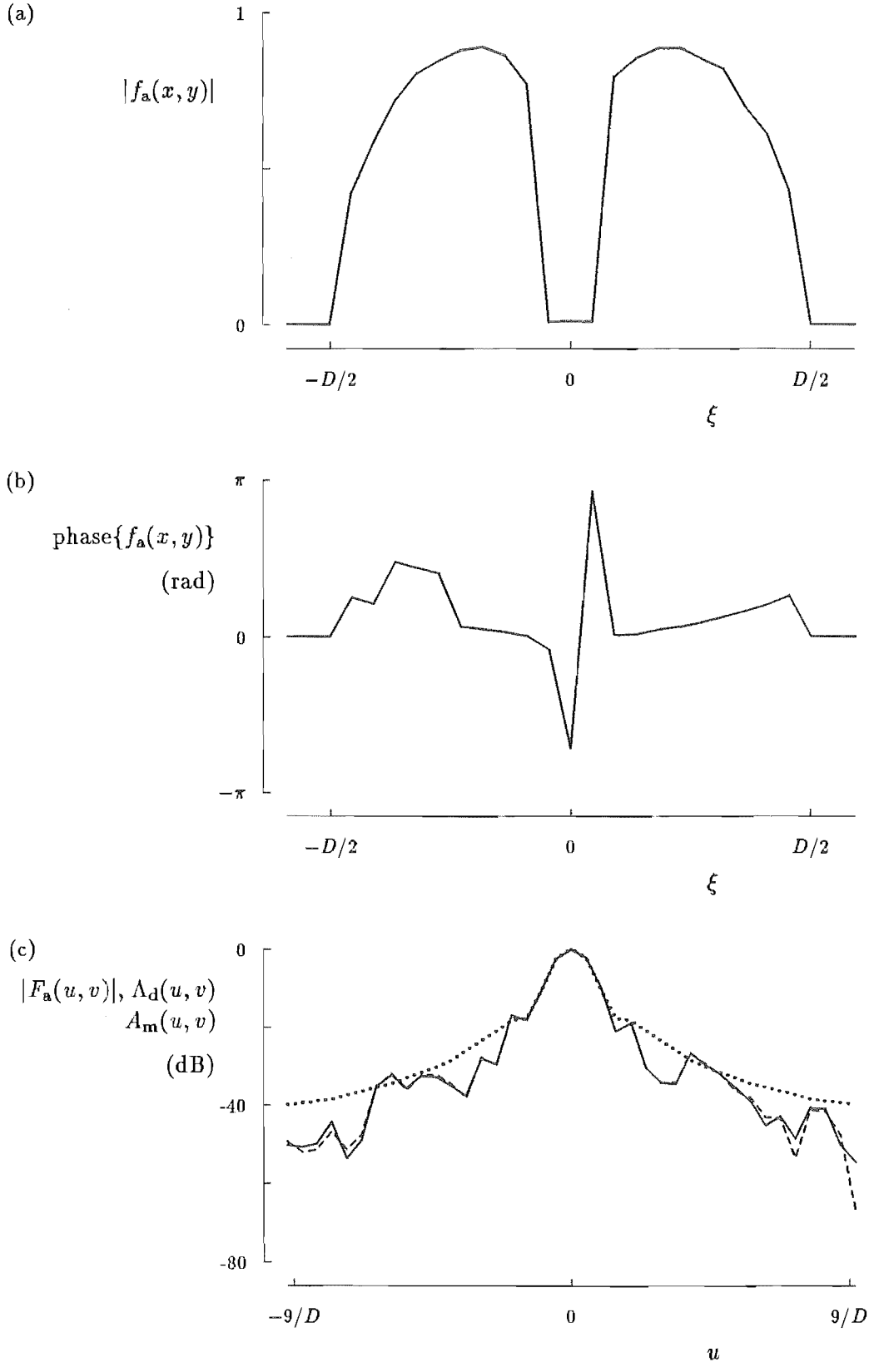


Figure 4.18 Actual copolar fields for the worked example: (a) actual copolar aperture field amplitude distribution; (b) actual copolar aperture field phase distribution; (c) actual copolar far field amplitude pattern (solid curve), measured copolar far field amplitude pattern (dashed curve) and design envelope (dotted curve). The variable ξ, which identifies position along a diagonal in the aperture plane, is defined in (4.8).

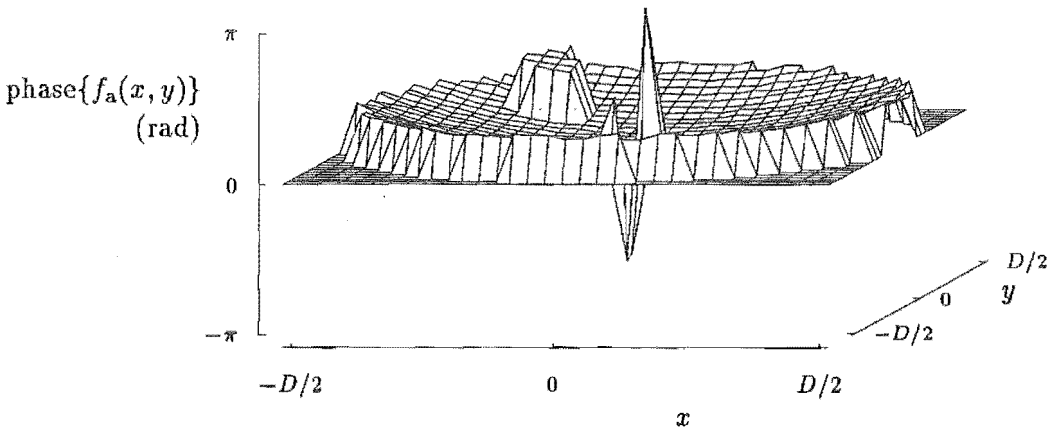


Figure 4.19 Actual copolar aperture field phase distribution for the worked example.

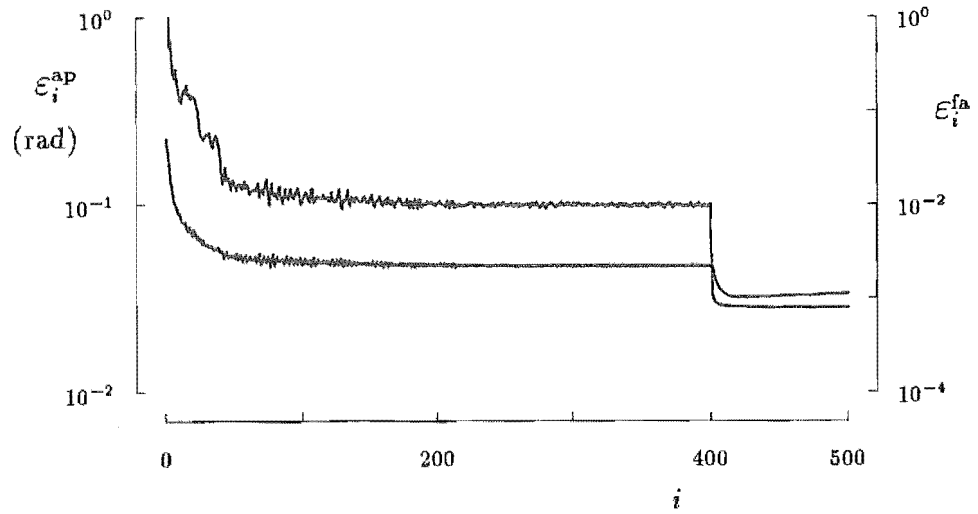


Figure 4.20 Error curves for the worked example. The curves are of $\mathcal{E}_i^{fa}$ (lower curve) and $\mathcal{E}_i^{ap}$ (upper curve) and were generated by the CC algorithm run which was chosen in step (3) of the composite algorithm (Sec. 4.4.5).

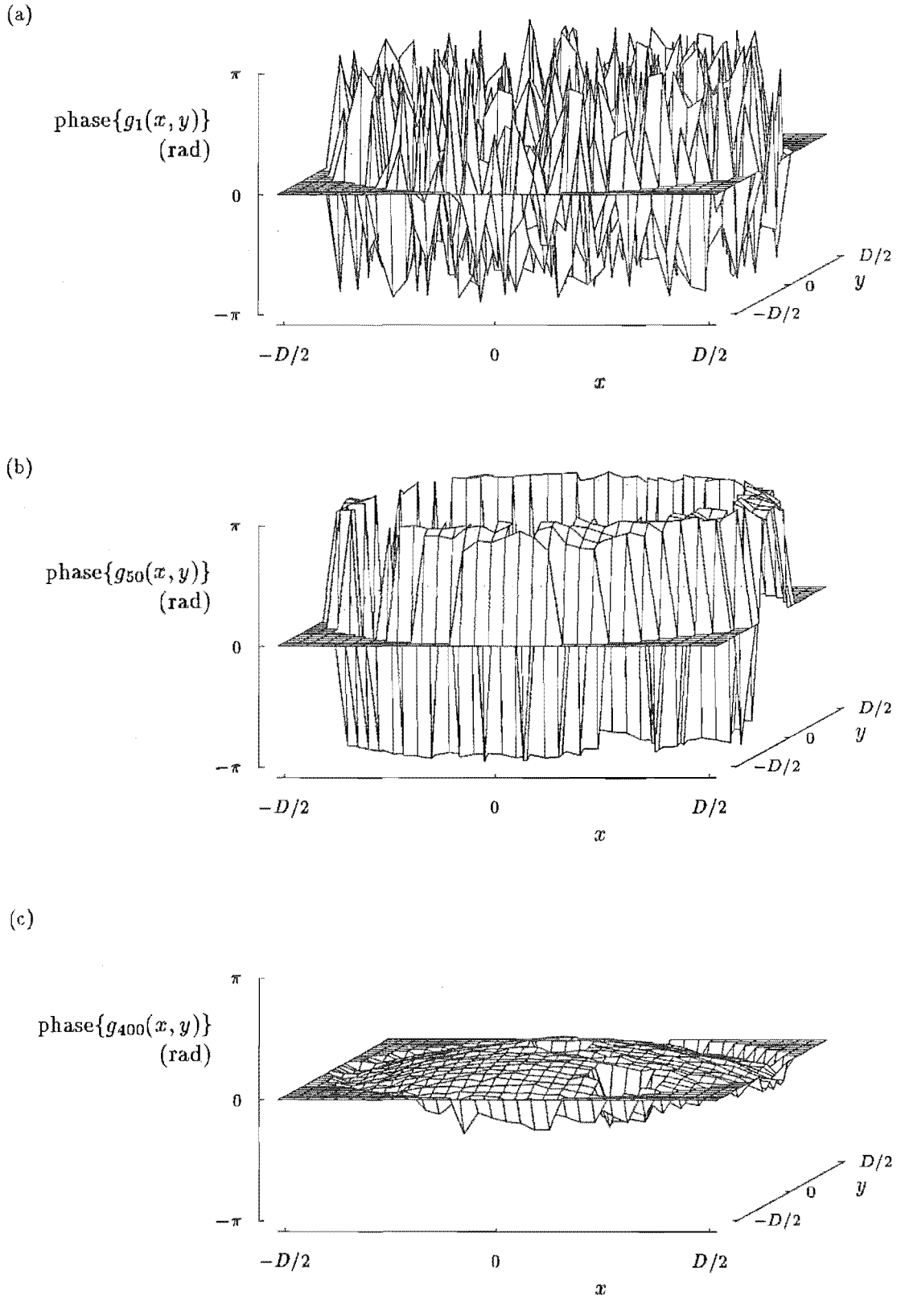
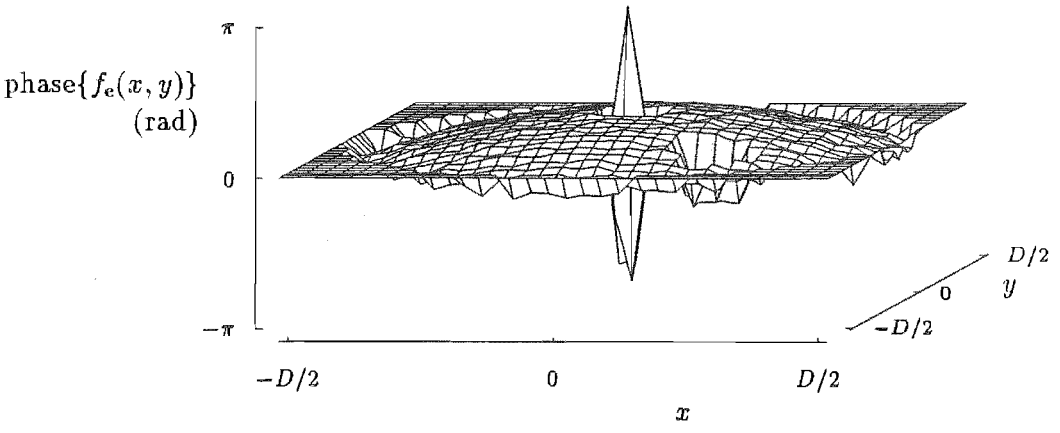
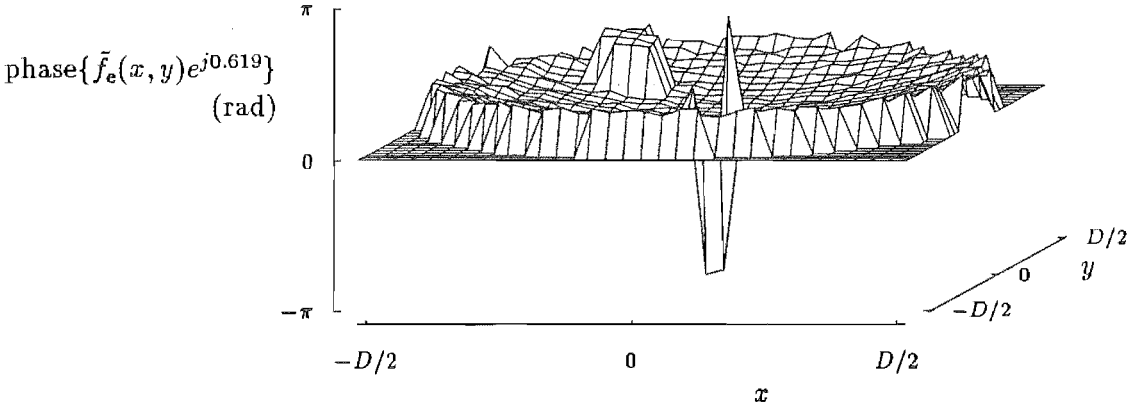


Figure 4.21 Plots of $g_i(x, y)$ for the worked example: (a) $i = 1$; (b) $i = 50$; (c) $i = 400$. The distribution for $i = 500$ is plotted in Figure 4.22(a). All except the first of these distributions were generated by the CC algorithm.

(a)



(b)



(c)

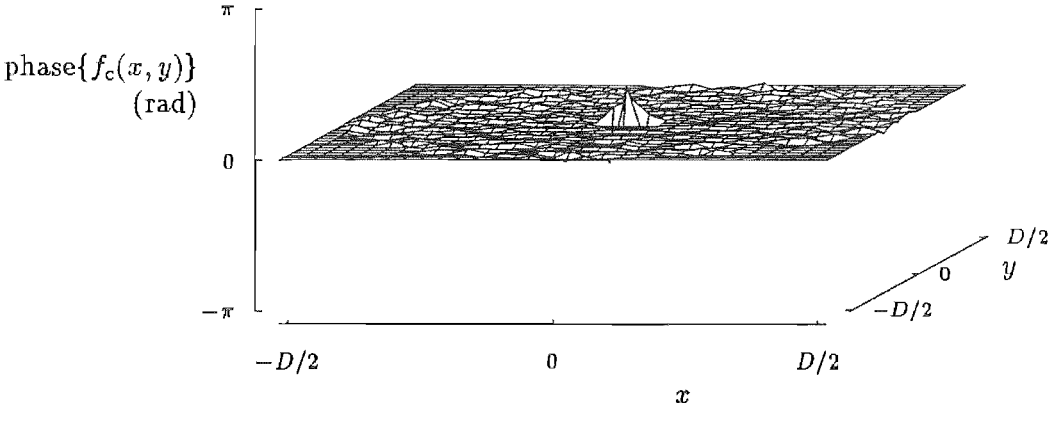


Figure 4.22 Plots of the estimated and corrected copolar aperture field phase distributions for the worked example: (a) estimated copolar aperture field phase distribution generated by the composite algorithm; (b) phase distribution of the conjugate image of the estimated copolar aperture field distribution; (c) corrected copolar aperture field phase distribution.

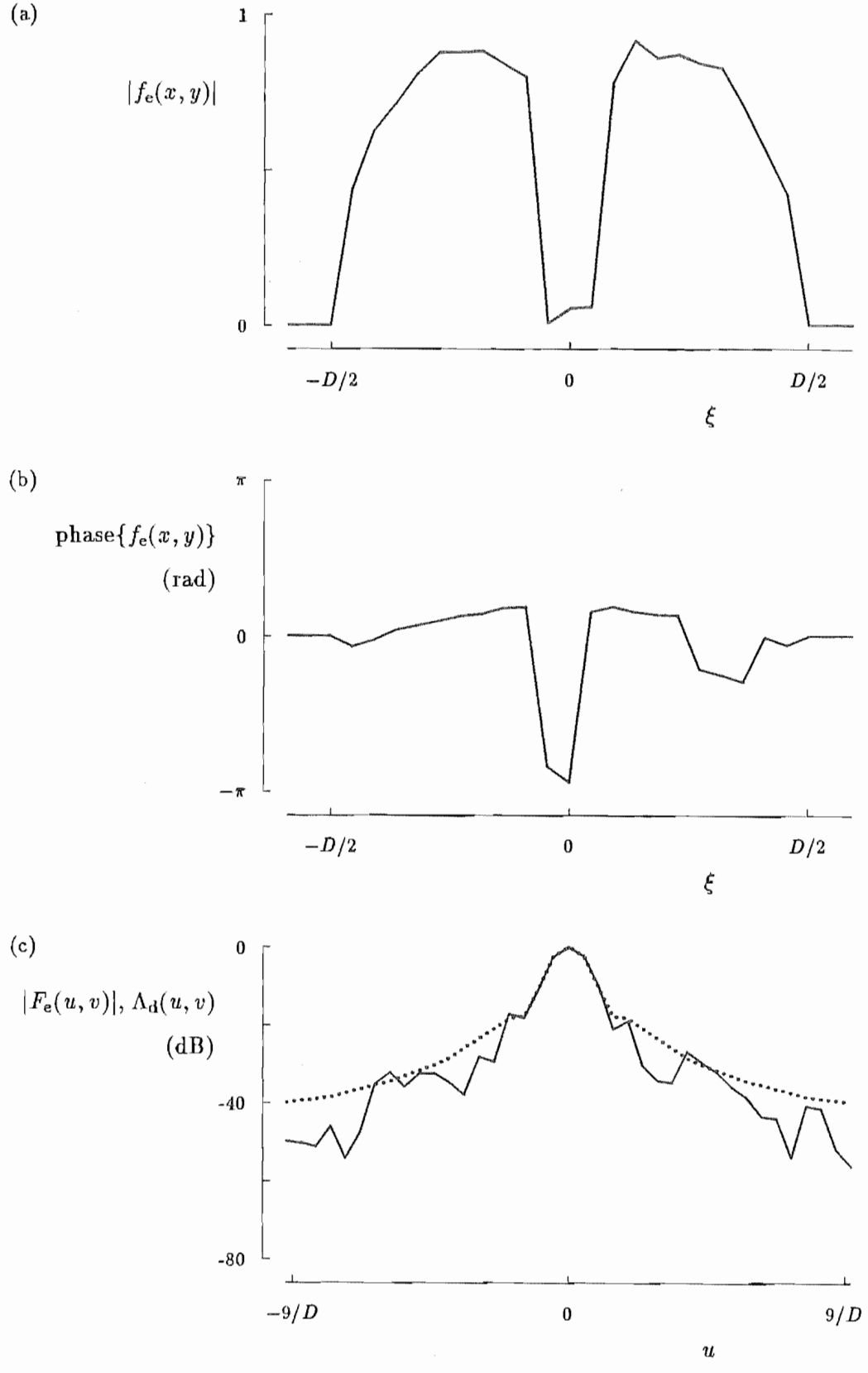


Figure 4.23 Estimated copolar fields for the worked example: (a) estimated copolar aperture field amplitude distribution; (b) estimated copolar aperture field phase distribution; (c) estimated copolar far field amplitude pattern (solid curve) and design envelope (dotted curve). The variable ξ, which identifies position along a diagonal in the aperture plane, is defined in (4.8).

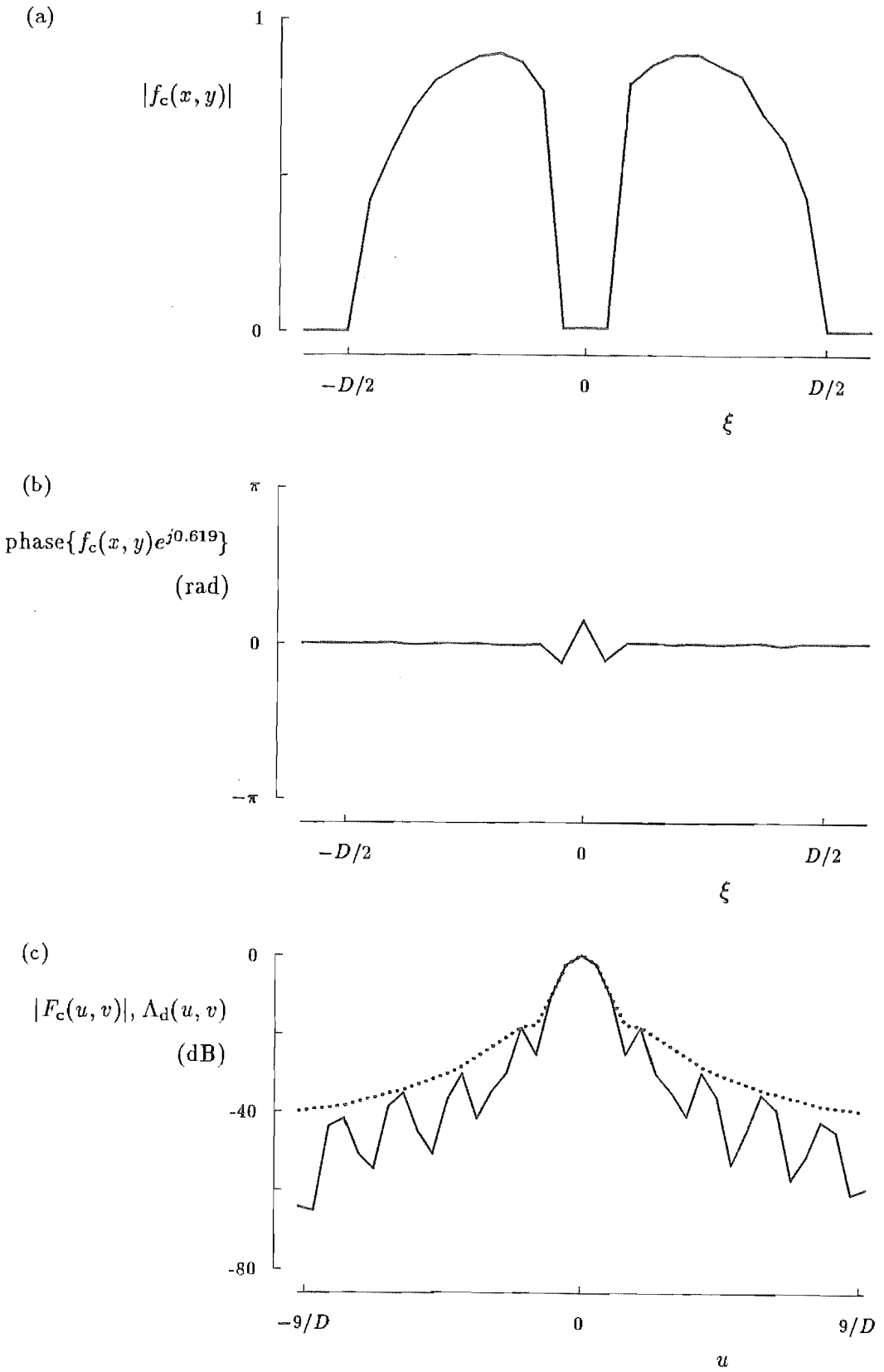


Figure 4.24 Corrected copolar fields for the worked example: (a) corrected copolar aperture field amplitude distribution; (b) corrected copolar aperture field phase distribution; (c) corrected copolar far field amplitude pattern (solid curve) and design envelope (dotted curve). The variable ξ, which identifies position along a diagonal in the aperture plane, is defined in (4.8).

4.6 RELATIONSHIPS BETWEEN ERROR MEASURES

This section discusses the relationship between $\mathcal{E}^{\mathrm{fa}}$, $\mathcal{E}^{\mathrm{ap}}$ and E_{c}, which are all defined in Section 4.3, for the CC and HIO algorithms. This relationship is important because the best of six CC and HIO algorithm runs is chosen, in step (3) of the composite algorithm (Sec. 4.4.5), on the basis of $\mathcal{E}^{\mathrm{fa}}$. However, as intimated in Section 4.3, $\mathcal{E}^{\mathrm{ap}}$ is a more direct indicator of how accurately the phase of $f_{\mathrm{e}}(x, y)$, generated by the CC and HIO algorithms, approximates the phase of the image-form of $f_{\mathrm{a}}(x, y)$ while E_{c} indicates the success of the correction process based upon the information contained in phase$\{f_{\mathrm{e}}(x, y)\}$.

Figure 4.26(a) shows a graph of $\mathcal{E}^{\mathrm{ap}}$ versus $\mathcal{E}^{\mathrm{fa}}$ for the CC and HIO algorithms applied to the basic model, which is defined by (4.47). Each point on the graph represents the value of $\mathcal{E}^{\mathrm{ap}}$ plotted against the corresponding value of $\mathcal{E}^{\mathrm{fa}}$, for the estimate $f_{\mathrm{e}}(x, y)$ of the image-form of $f_{\mathrm{a}}(x, y)$ generated by one run of either the CC algorithm or the HIO algorithm. The ellipses shown in Figure 4.26(a) correspond to 20 runs of the CC algorithm while the crosses correspond to 20 runs of the HIO algorithm. Each run was started with a different random phase distribution. Utilizing the results of these same runs, Figures 4.26(b) and (c) depict graphs of E_{c} versus $\mathcal{E}^{\mathrm{fa}}$ and $\mathcal{E}^{\mathrm{ap}}$ respectively. It should be kept in mind that when running the composite algorithm the CC and HIO algorithms are run only 3 times each. The latter algorithms were in fact run 20 times each to generate the graphs, shown in Figure 4.26, so as to make the trends in the relationships between the errors more easily discernible.

Figure 4.26(a) and (b) indicate that the smaller values of $\mathcal{E}^{\mathrm{fa}}$ correspond to the smaller values of both $\mathcal{E}^{\mathrm{ap}}$ and E_{c}. This is fortunate, because the composite algorithm chooses the best of 3 CC algorithm runs and 3 HIO algorithm runs on the basis of $\mathcal{E}^{\mathrm{fa}}$. Nevertheless, had it been possible to make the choice on the basis of $\mathcal{E}^{\mathrm{ap}}$ or E_{c} it is likely that the same run would have been chosen. Or, putting it another way, it is reasonably certain that the CC or HIO algorithm run with the smallest value $\mathcal{E}^{\mathrm{fa}}$ also has the smallest values of $\mathcal{E}^{\mathrm{ap}}$ and E_{c}. Note, however, that a relatively large value of

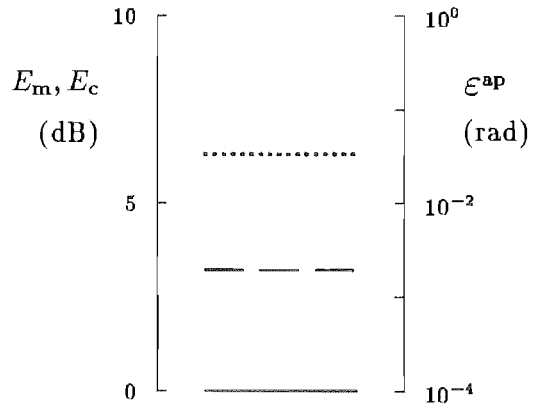


Figure 4.25 Errors for the worked example. The values of $\mathcal{E}^{\mathrm{ap}}$ (dotted curve), E_{m} (dashed curve) and E_{c} (solid curve) pertain to the composite algorithm applied to the basic model.

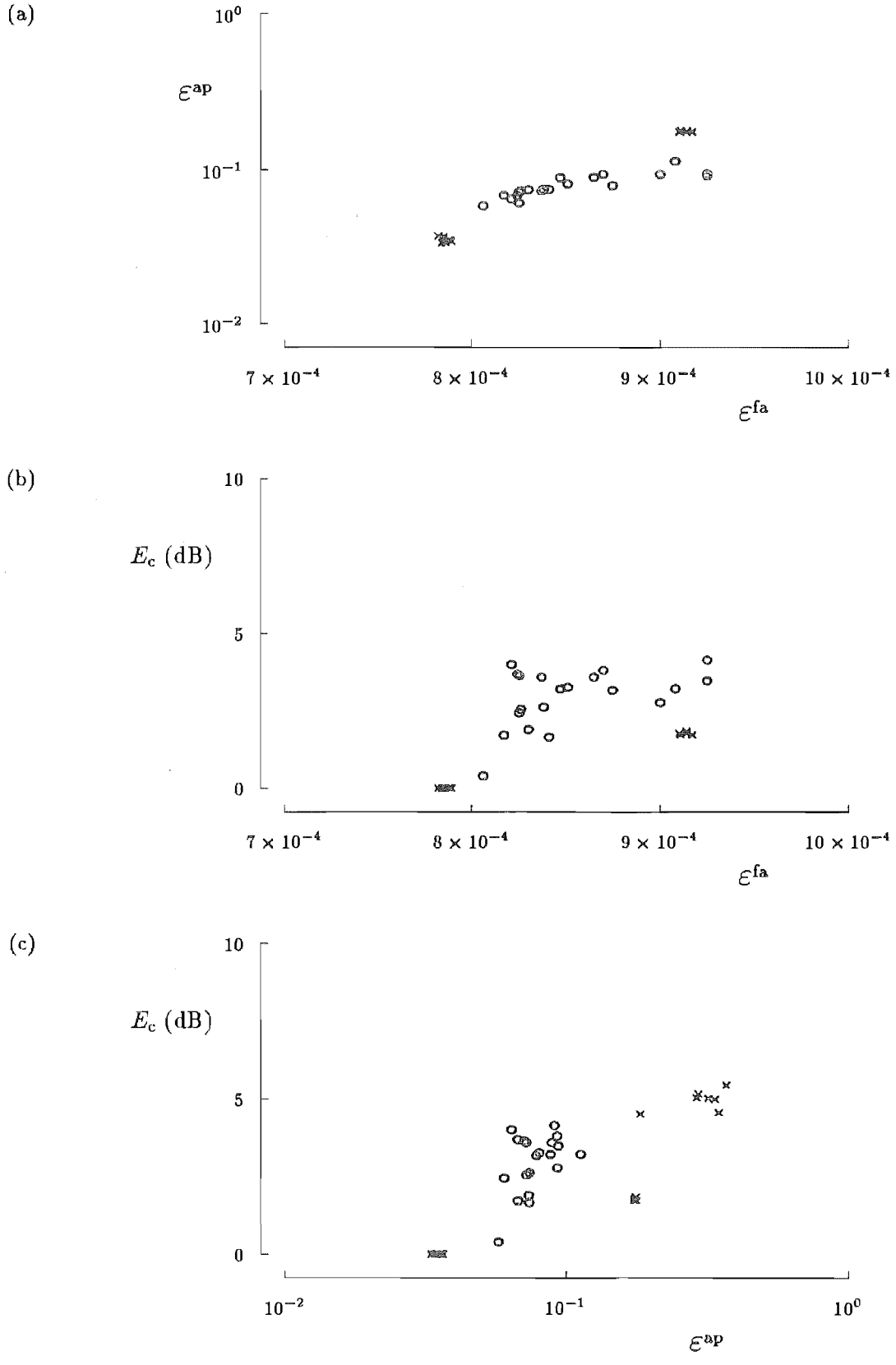


Figure 4.26 Relationship between errors $\mathcal{E}^{ap}$, $\mathcal{E}^{fa}$ and E_c, for the CC and HIO algorithms applied to the basic model: (a) $\mathcal{E}^{ap}$ versus $\mathcal{E}^{fa}$; (b) E_c versus $\mathcal{E}^{fa}$; (c) E_c versus $\mathcal{E}^{ap}$. The ellipses correspond to 20 runs of the CC algorithm and the crosses correspond to 20 runs of the HIO algorithm.

$\mathcal{E}^{\mathrm{fa}}$ or $\mathcal{E}^{\mathrm{ap}}$ does not necessarily imply a relatively large value of E_{c}, as can be seen in Figures 4.26(b) and (c).

Considering the clear trends shown in the graphs in Figure 4.26, it might be thought that a better strategy, than that on which the composite algorithm is based, would be to set a threshold value for $\mathcal{E}^{\mathrm{fa}}$ and to run the CC (or HIO) algorithm for as many times as required to find a run for which $\mathcal{E}^{\mathrm{fa}}$ is less than the threshold. The result of that run would then be chosen to be $f_{\mathrm{c}}(x, y)$. However, there does not appear to be any clear criterion for choosing an appropriate threshold for $\mathcal{E}^{\mathrm{fa}}$. Consider the results shown in Figure 4.26. The threshold would have to lie somewhere between the minimum and maximum values of $\mathcal{E}^{\mathrm{fa}}$, which are 7.82×10^{-4} and 9.25×10^{-4} respectively. Note that the maximum value is only 20% larger than the minimum value. My experience of the modified Gerchberg-Saxton algorithm applied to many particular models suggest that such a small range of values for $\mathcal{E}^{\mathrm{fa}}$ is typical. Therefore, in order to set a meaningful threshold, the minimum value of $\mathcal{E}^{\mathrm{fa}}$ needs to be known to an accuracy of significantly better than 20%. The minimum value of $\mathcal{E}^{\mathrm{fa}}$ depends upon the measurement inaccuracies and is typically proportional to Γ_{ran}. However, should the algorithm be applied in real-world situations, the noise level inherent in the measurements is unlikely to be known to a relative accuracy of better than 20%. Therefore, one cannot expect to predict an accurate enough estimate for the minimum value of $\mathcal{E}^{\mathrm{fa}}$. This implies that it is impracticable to choose a meaningful threshold value for $\mathcal{E}^{\mathrm{fa}}$. Note that the situation is not improved if $\mathcal{E}^{\mathrm{fa}}$ is replaced with the image error $\mathcal{E}^{\mathrm{I}}$, defined in (3.71), because in my experience is that $\mathcal{E}^{\mathrm{I}}$ is almost proportional to $\mathcal{E}^{\mathrm{fa}}$.

4.7 FAR FIELD MEASUREMENT CONSIDERATIONS

It is pointed out in Section 3.3.4 that, if the time taken to perform a measurement is appreciable, this measurement time can account for a significant fraction of the overall measurement cost. In general, the more samples of the copolar far field amplitude pattern that are required, the longer is the measurement time. It could therefore be economically advantageous to reduce, in so far as is feasible, the number of measured samples of the copolar far field amplitude pattern. This can be achieved by increasing the spacing between the sample points or by reducing the area of the region in the u, v plane over which the amplitude pattern is sampled. Another factor, which can affect the expense involved in measuring the copolar far field amplitude pattern, is the accuracy required. A more accurate measurement often requires more care and more expensive equipment.

Section 4.7.1 describes an algorithm which preprocesses the measured data in an attempt to remove some of its imperfections. This algorithm, and its effect on the convergence properties of the composite algorithm, are illustrated by computational example. Section 4.7.2 discusses and illustrates the effect, on the convergence properties of the composite algorithm, of altering what is here called the far field sampling factor. Three methods for enabling the composite algorithm to be applied to truncated far field data are described, discussed and illustrated with the aid of an example in Section 4.7.3.

4.7.1 Smoothing far field data

In this section, what is here called the smoothing algorithm is described. The purpose of this algorithm is to remove some of the imperfections inevitably present in samples

of $A_m(u,v)$. It is a preprocessing algorithm in the sense that it operates on the samples of $A_m(u,v)$ to generate massaged samples of $A_m(u,v)$, which can then be fed to the modified Gerchberg-Saxton algorithm.

A way of determining the accuracy of the measured copolar far field amplitude pattern $A_m(u,v)$ is to inspect what is here called the *approximate autocorrelation* $ff_m(x,y)$ generated by $A_m(u,v)$. The approximate autocorrelation is defined to be

$$ff_m(x,y) = \mathrm{IFT}\Big\{[A_m(u,v)]^2\Big\} \tag{4.48}$$

It follows from the autocorrelation theorem (3.42) that if $A_m(u,v) = |F_a(u,v)|$ then $ff_m(x,y)$ is equal to the autocorrelation $ff_a(x,y)$ of $f_a(x,y)$. Because $f_a(x,y)$ is compact, as defined in Section 3.4.1.1, it follows from (3.43) that $ff_a(x,y)$ is also compact. Therefore $ff_a(x,y)$ vanishes outside its support, here denoted by S^{auto}. However, the inevitable measurement inaccuracies imply that $ff_m(x,y) \neq ff_a(x,y)$. In particular, $ff_m(x,y)$ tends to have non-zero values outside S^{auto}. This is illustrated in Figure 4.27 which shows cuts through the approximate autocorrelations corresponding to the measured copolar far field amplitude patterns depicted in Figure 4.10. In general, the more inaccurate $A_m(u,v)$ is, the higher the relative levels of $ff_m(x,y)$ are outside S^{auto}. Note that, because the supports of $f_a(x,y)$ and $f_d(x,y)$ are, in general, equal, S^{auto} can be straightforwardly computed, even when $f_a(x,y)$ is unknown, by equating it to the support of the autocorrelation of $f_d(x,y)$.

Because $ff_a(x,y)$ is compact, its Fourier transform $|F_a(u,v)|^2$ varies smoothly from sample to sample. On the other hand, measurement noise tends to make $[A_m(u,v)]^2$ vary erratically from sample to sample. This suggests a way of removing some of the imperfection present in $A_m(u,v)$. The data $[A_m(u,v)]^2$ can be smoothed by applying the basic iterative Fourier transform algorithm (Sec. 3.4.3), with the image information consisting of S^{auto} and the Fourier information being that $[A_m(u,v)]^2$ must be real and non-negative. Accordingly, the $i_{\mathrm{pre}}^{\mathrm{th}}$ iteration of what is here called the *smoothing algorithm* is defined by (cf. Fig. 3.9)

$$\begin{aligned}
H_{i_{\mathrm{pre}}}(u,v) &= \mathrm{FT}\{h_{i_{\mathrm{pre}}}(x,y)\} \\
H'_{i_{\mathrm{pre}}}(u,v) &= \begin{cases} H_{i_{\mathrm{pre}}}(u,v) & \text{if } H_{i_{\mathrm{pre}}}(u,v) \geq 0 \\ 0 & \text{otherwise} \end{cases} \\
h'_{i_{\mathrm{pre}}}(x,y) &= \mathrm{IFT}\Big\{H'_{i_{\mathrm{pre}}}(u,v)\Big\} \\
h_{i_{\mathrm{pre}}+1}(x,y) &= \begin{cases} h'_{i_{\mathrm{pre}}}(x,y) & \text{for } (x,y) \in S^{\mathrm{auto}} \\ 0 & \text{elsewhere} \end{cases}
\end{aligned} \tag{4.49}$$

where $h_{i_{\mathrm{pre}}}(x,y)$ is an estimate of $ff_a(x,y)$. The algorithm is started by setting $h_1(x,y) = ff_m(x,y)$. If the algorithm is run for I_{pre} iterations, $H'_{I_{\mathrm{pre}}}(u,v)$ is taken to be a smoothed version of $[A_m(u,v)]^2$. This algorithm is to be thought of as a preliminary to the composite algorithm, so that the occurrences of $A_m(u,v)$, in the composite algorithm, can be replaced by the positive square root of $H'_{I_{\mathrm{pre}}}(u,v)$. Note that running the smoothing algorithm for merely one iteration is equivalent to not running the algorithm at all, because $[H'_1(u,v)]^{1/2} = A_m(u,v)$

The smoothness of $H'_{i_{\mathrm{pre}}}(u,v)$ can be monitored by the *autocorrelation error* $\mathcal{E}^{\mathrm{auto}}_{i_{\mathrm{pre}}}$

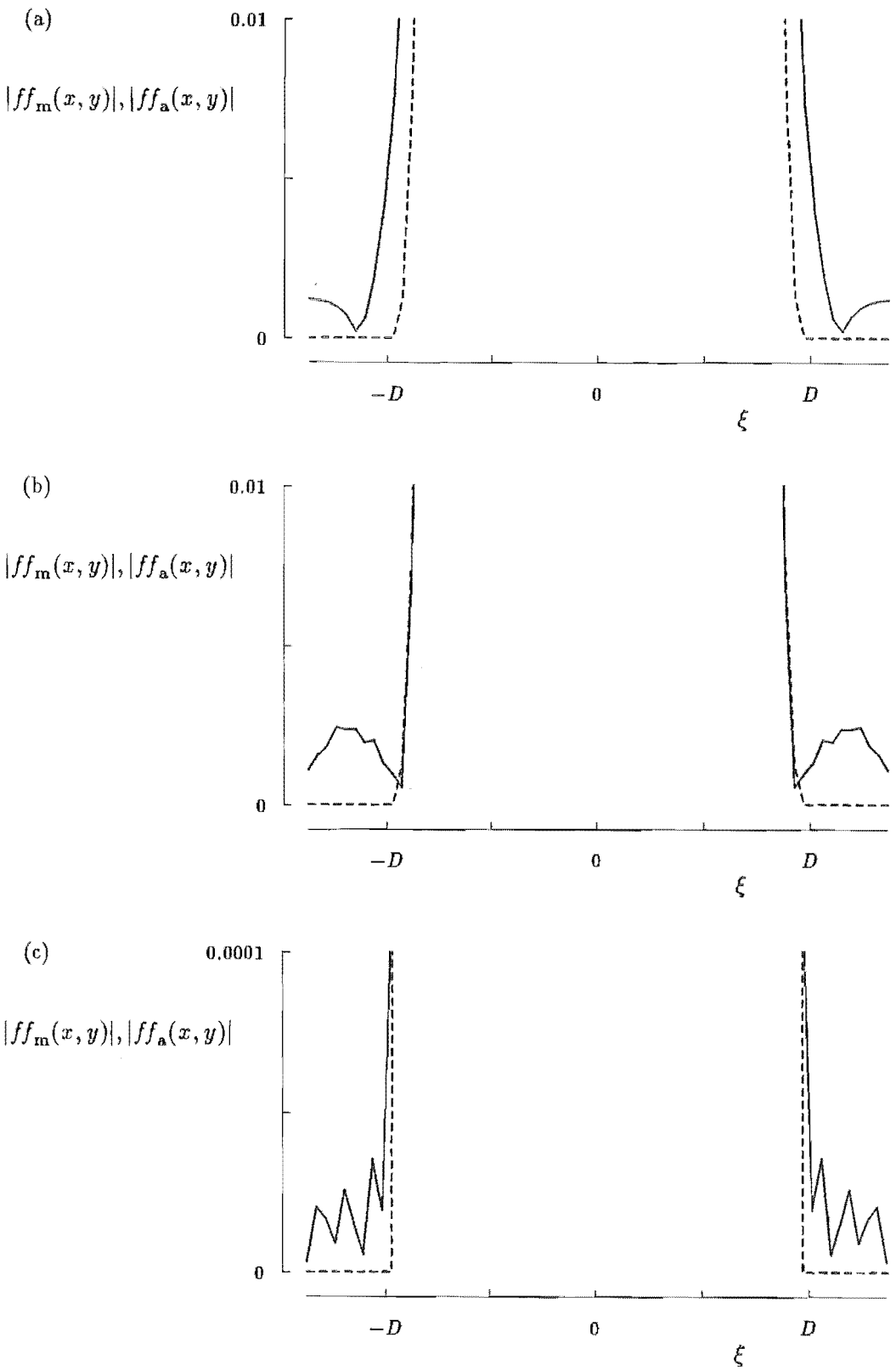


Figure 4.27 Actual and approximate autocorrelations corresponding to different measurement inaccuracies: (a) $\Gamma_{\mathrm{cal}} = 1.05$; (b) $\Gamma_{\mathrm{ran}} = -50$ dB; (c) $D^{A_{\mathrm{m}}} = 15/D$. In each case design 2 is employed. The amplitudes of $ff_{\mathrm{m}}(x,y)$ and $ff_{\mathrm{a}}(x,y)$ are represented by solid curves and dashed curves respectively and are normalized to have peak values of unity. The variable ξ, which identifies position along a diagonal in the aperture plane, is defined in (4.6). Note that the several $ff_{\mathrm{m}}(x,y)$ depicted in this figure are derived, via (4.48), from the several $A_{\mathrm{m}}(u,v)$ depicted in Figure 4.10.

which is defined to be

$$\mathcal{E}^{\text{auto}}_{i_{\text{pre}}} = \frac{1}{h'_{i_{\text{pre}}}(0,0)} \left[\frac{\iint_{(x,y)\notin S^{\text{auto}}} |h'_{i_{\text{pre}}}(x,y)|^2 \, dx \, dy}{\iint_{(x,y)\notin S^{\text{auto}}} dx \, dy} \right]^{1/2} \tag{4.50}$$

which is the rms value of $h'_{i_{\text{pre}}}(x,y)$ outside S^{auto}, normalized to its maximum value $h'_{i_{\text{pre}}}(0,0)$. A measure of how accurately the square root of $H'_{i_{\text{pre}}}(u,v)$ approximates the actual copolar far field amplitude pattern is given by the *far field data error* $\bar{\mathcal{E}}^{\text{fa}}_{i_{\text{pre}}}$, which is defined by

$$\bar{\mathcal{E}}^{\text{fa}}_{i_{\text{pre}}} = \frac{1}{|F_{\text{a}}(0,0)|} \left[\frac{\iint \left[[H'_{i_{\text{pre}}}(u,v)]^{1/2} - |F_{\text{a}}(u,v)| \right]^2 \, du \, dv}{\iint du \, dv} \right]^{1/2} \tag{4.51}$$

Note that $\bar{\mathcal{E}}^{\text{fa}}_{i_{\text{pre}}}$ differs from $\mathcal{E}^{\text{fa}}$ because the latter, which is defined by (4.23), indicates the difference between $|F_{\text{e}}(u,v)|$ and $A_{\text{m}}(u,v)$. The reason for introducing $\bar{\mathcal{E}}^{\text{fa}}_{i_{\text{pre}}}$ is that it indicates whether the smoothed version of $A_{\text{m}}(u,v)$ is more or less imperfect than the original version of $A_{\text{m}}(u,v)$.

Figure 4.28(a) depicts error curves of $\mathcal{E}^{\text{auto}}_{i_{\text{pre}}}$ and $\bar{\mathcal{E}}^{\text{fa}}_{i_{\text{pre}}}$ for the smoothing algorithm applied to data obtained from the basic model. After only ten iterations, $\mathcal{E}^{\text{auto}}_{i_{\text{pre}}}$ has reduced drastically to about one thousandth of its initial value. By contrast, after one iteration all values of $\bar{\mathcal{E}}^{\text{fa}}_{i_{\text{pre}}}$ are 28% larger than $\bar{\mathcal{E}}^{\text{fa}}_{1}$. This is encouraging, because the smoothing algorithm does not incorporate a constraint to force $[H'_{i_{\text{pre}}}(u,v)]^{1/2}$ to be approximately equal to either $|F_{\text{a}}(u,v)|$ or $A_{\text{m}}(u,v)$. Yet the values of $\bar{\mathcal{E}}^{\text{fa}}_{i_{\text{pre}}}$ indicate that amount by which $[H'_{i_{\text{pre}}}(u,v)]^{1/2}$ differs from $|F_{\text{a}}(u,v)|$ is only 28% greater than the amount by which $A_{\text{m}}(u,v)$ differs from $|F_{\text{a}}(u,v)|$.

The effect on the composite algorithm of applying it to a smoothed $A_{\text{m}}(u,v)$ is illustrated in Figure 4.28(b). To obtain the data for this figure, the smoothing algorithm was run for I_{pre} iterations. Then $A_{\text{m}}(u,v)$ was replaced by $[H'_{I_{\text{pre}}}(u,v)]^{1/2}$ and the composite algorithm was run. The errors $\mathcal{E}^{\text{ap}}$, E_{m} and E_{c} were then computed, where E_{m} is the envelope error (Sec. 4.3) for the smoothed data. This procedure was repeated many times, using a different value of I_{pre} each time, so that the errors could be graphed against I_{pre} in Figure 4.28(b). It is seen from the graph in Figure 4.28(b) that, unfortunately, the smoothing of $A_{\text{m}}(u,v)$ hinders, rather than improves, the convergence of the composite algorithm. For this reason the smoothing algorithm is not invoked throughout the remainder of this thesis.

4.7.2 Need for oversampling the far field

This section discusses the consequences of altering the factor by which $A_{\text{m}}(u,v)$ is oversampled. The question of the uniqueness of a solution is addressed and the effect of the sampling factor (as defined in Sec. 3.4.1.3) on the convergence properties of the composite algorithm is described.

In the terminology utilized throughout this chapter, a 'solution' to the Fourier phase problem (3.41) is here said to be any copolar aperture field distribution $f_{\text{e}}(x,y)$ which is approximately zero outside the aperture support S^{aper} and whose amplitude pattern $|F_{\text{e}}(u,v)|$ approximates $A_{\text{m}}(u,v)$ to within a value of $\mathcal{E}^{\text{fa}}$ which is less than

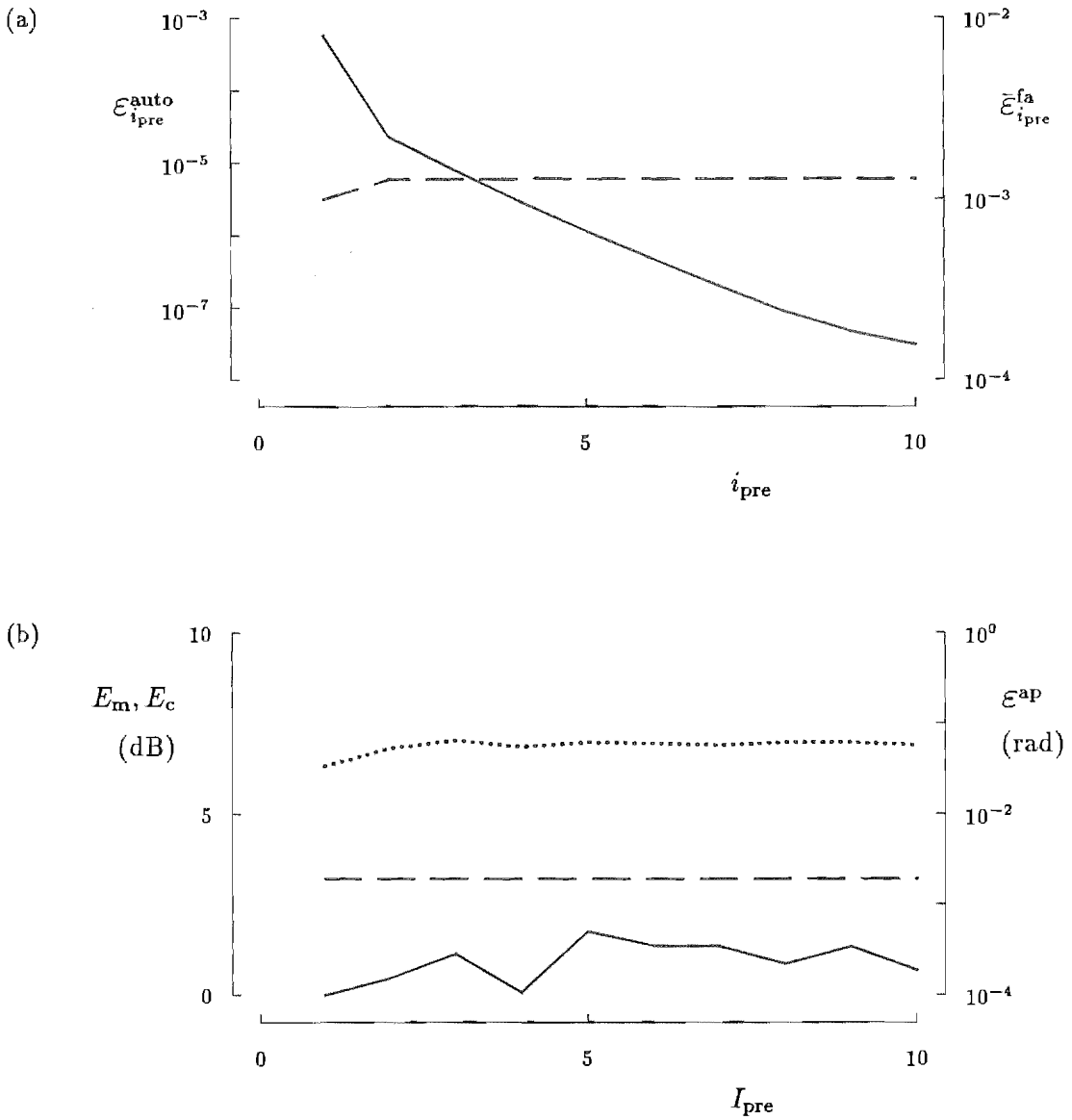


Figure 4.28 Effect of the smoothing algorithm on the composite algorithm: (a) error curves of the autocorrelation error $\mathcal{E}^{\text{auto}}_{i_{\text{pre}}}$ (solid curve) and far field data error $\bar{\mathcal{E}}^{\text{fa}}_{i_{\text{pre}}}$ (dashed curve) for the smoothing algorithm; (b) values of $\mathcal{E}^{\text{ap}}$ (dotted curve), E_{m} (dashed curve) and E_{c} (solid curve) pertaining to the application of the composite algorithm after I_{pre} iterations of the smoothing algorithm. The algorithm is applied to data generated from the basic model.

the measurement noise Γ_{ran}. A solution is said to be exact if $|F_{\text{e}}(u,v)| = A_{\text{m}}(u,v)$. A solution is here said to be correct if it approximates the image-form of $f_{\text{a}}(x,y)$ accurately enough for the value of E_{c} to equal 0 dB. All other solutions are said to be incorrect. If a solution is unique then it must also be correct, because $f_{\text{e}}(x,y) = f_{\text{a}}(x,y)$ is always a solution.

Section 3.4.2 explains that, in order to obtain a unique solution to the Fourier phase problem, it is necessary to oversample $A_{\text{m}}(u,v)$ by a factor of at least two. As intimated in Section 3.4.2.1, the requirement to oversample $A_{\text{m}}(u,v)$ by a factor of at least two seems reasonable because, only when this requirement is met, can the continuous actual copolar far field amplitude pattern be estimated throughout the region of the u,v plane occupied by the measured samples.

However, the particular radio engineering phase problem which is solved by the composite algorithm is somewhat different from the Fourier phase problem. In the Fourier phase problem, the only available information is $A_{\text{m}}(u,v)$ (see (3.41)). Provided this is oversampled by a factor of at least two, the extents of $f_{\text{a}}(x,y)$ can be estimated from $A_{\text{m}}(u,v)$ by invoking (3.42) and (3.43). On the other hand, the composite algorithm has additional information availiable to it. The size and shape of the aperture and, therefore, the support S^{aper} of $f_{\text{a}}(x,y)$ are usually available, as intimated in Section 4.1.1. Furthermore, $|f_{\text{d}}(x,y)|$, which is an estimate of $|f_{\text{a}}(x,y)|$, is also known. This additional information could well imply that the composite algorithm might generate a unique solution when operating on far field data which is oversampled by a factor of less than two.

To simulate the oversampling of $A_{\text{m}}(u,v)$ by various factors, the computer model described in Section 4.2 is changed by replacing (4.4) with the following equation (cf. (3.44) and the first equation of (3.37)):

$$D = \frac{64\Delta_x}{\alpha_u} = \frac{64\Delta_y}{\alpha_v} \tag{4.52}$$

where α_u and α_v, which are the sampling factors in the u and v directions respectively, are set to equal each other.

Figure 4.29 depicts results of 9 runs of the composite algorithm applied to data generated from the basic model. For each run, the sampling factors are set to different values in the range 1 to 4. For the different values of α_u and α_v, the error measures $\mathcal{E}^{\text{ap}}$, E_{m} and E_{c} are plotted in Figure 4.29(a) and the far field amplitude error $\mathcal{E}^{\text{fa}}$ is plotted in Figure 4.29(b). Note that the value of E_{m} is different for each value of α_u and α_v. This is because, from its definition (4.24), E_{m} is the maximum envelope error taken over all sample points in the u,v plane. For different values of α_u and α_v, the sample points are differently located in the u,v plane. Since the envelope error varies continuously over the u,v plane, its maximum value is likely to be different when calculated over different sets of points in the plane. This is a manifestation of the picket fence effect, which is explained in Section 4.2.1.

It is apparent from the behaviour of E_{c} in Figure 4.29(a) that $f_{\text{e}}(x,y)$, generated by the composite algorithm, is a correct solution whenever $A_{\text{m}}(u,v)$ is oversampled by a factor of at least 1.7.

The $\mathcal{E}^{\text{fa}}$ curve in Figure 4.29(b) indicates that when $\alpha_u = \alpha_v = 1$, the generated $f_{\text{e}}(x,y)$ is an exact solution to the Fourier phase problem. Yet, the corresponding values of $\mathcal{E}^{\text{ap}}$ and E_{c}, shown in Figure 4.29(a), indicate that this exact solution is an incorrect solution. When $\alpha_u = \alpha_v = 1.2$ and 1.4, the solutions are neither correct nor exact but have smaller values of $\mathcal{E}^{\text{fa}}$ than do the solutions obtained for larger sampling

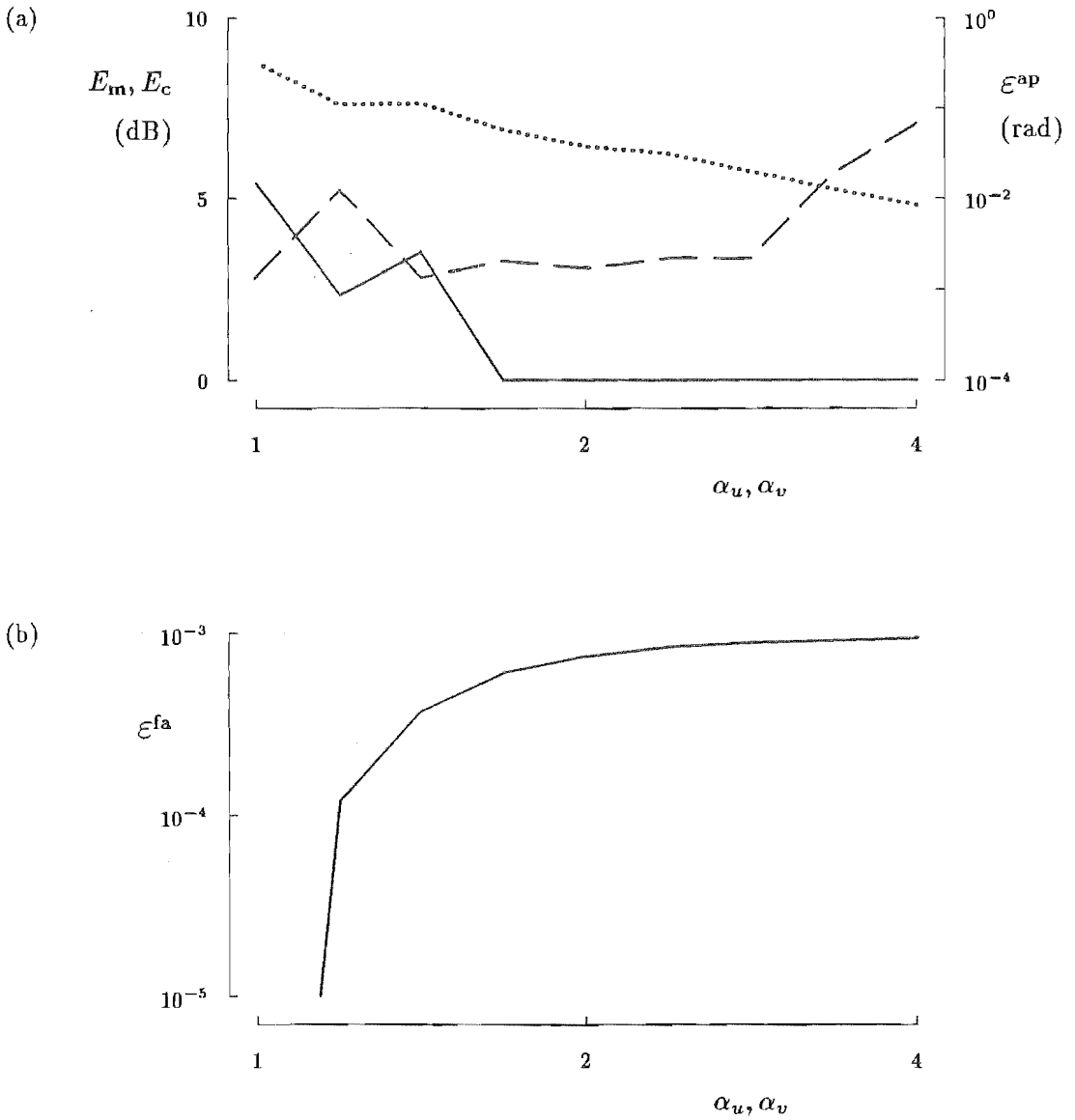


Figure 4.29 Effect of sampling factor on the convergence of the composite algorithm: (a) values of $\mathcal{E}^{ap}$ (dotted curve), E_m (dashed curve) and E_c (solid curve) plotted against far field sampling factors α_u and α_v; (b) values of far field amplitude error $\mathcal{E}^{fa}$ plotted against the same sampling factors. The errors pertain to the composite algorithm applied to data from the basic model which has been modified to accommodate various values of $\alpha_u = \alpha_v$.

factors. This behaviour implies that, for the composite algorithm, the far field error $\mathcal{E}^{\mathrm{fa}}$ fails to indicate the accuracy to which $f_{\mathrm{e}}(x,y)$ approximates the image-form of $f_{\mathrm{a}}(x,y)$ when α_u and α_v are less than about 1.7. Furthermore, the solutions generated by the composite algorithm are not unique for these values of α_{u} and α_v. It is advisable, therefore, to ensure that α_u and α_v are both greater than or equal to about 1.7 when applying the composite algorithm.

For sampling factors of at least 1.7, the value of $\mathcal{E}^{\mathrm{ap}}$ decreases as the sampling factors increase. An intuitive explanation for this is now given. Consider the set of samples comprising $A_{\mathrm{m}}(u,v)$ oversampled by a factor of four. When there is no measurement noise, one quarter of these samples, corresponding to every second sample point in both the u and v directions, completely describes the amplitude of the continuous copolar far field radiation pattern at all points in the u,v plane. The remaining samples are therefore redundant. However, when $A_{\mathrm{m}}(u,v)$ includes measurement noise, which is independent from sample to sample, appropriate use of the redundant samples can be expected to reduce the effective level of the noise. This does indeed seem to be the case, because, as indicated by $\mathcal{E}^{\mathrm{ap}}$, the composite algorithm applied to data which is oversampled by a factor of four generates a more accurate solution $f_{\mathrm{e}}(x,y)$ than when it is applied to data oversampled by a factor of two.

As intimated earlier in this section a solution to the Fourier phase problem is unique when $A_{\mathrm{m}}(u,v)$ is oversampled by a factor of at least two. This implies that a solution, generated by the composite algorithm, is guaranteed to be unique when $A_{\mathrm{m}}(u,v)$ is oversampled by a factor of at least two. However, there is no such guarantee when $A_{\mathrm{m}}(u,v)$ is oversampled by a factor of less than two. To avoid converging to incorrect solutions, the composite algorithm is always applied to data oversampled by a factor of at least two throughout the remainder of this thesis. However, it is worth keeping in mind that, in accord with results presented in this section, a smaller sampling factor, provided it is not less than about 1.7, may be satisfactory.

4.7.3 Truncated far field data

The model presented in Section 4.2 simulates truncated data by setting $A_{\mathrm{m}}(u,v)$ to zero outside a circular far field support denoted by $S^{A_{\mathrm{m}}}$. The diameter of this support is denoted by $D^{A_{\mathrm{m}}}$. Three different approaches are here explored for dealing with truncated far field data. In Section 4.7.3.1 the composite algorithm is applied directly to the truncated data. In Section 4.7.3.2 the data is extrapolated before being operated upon by the composite algorithm. The approach taken in Section 4.7.3.3 is based on an extention to the composite algorithm into which data extrapolation is incorporated. The approach taken in each of these sections is illustrated with a computational example. The approaches are compared in Section 4.7.3.4.

4.7.3.1 Direct application of the composite algorithm

Figure 4.30(a) shows the errors resulting from several runs of the composite algorithm, where each run was applied to far field data whose truncation is defined by a different value of $D^{A_{\mathrm{m}}}$. In all results presented in this section, $\mathcal{E}^{\mathrm{fa}}$ and E_{m} are computed only within $S^{A_{\mathrm{m}}}$, because it is only in this region that $A_{\mathrm{m}}(u,v)$ represents an approximation to $|F_{\mathrm{a}}(u,v)|$. However, E_{c} is computed over the whole u,v plane, because even if the measurements are confined to $S^{A_{\mathrm{m}}}$, the corrected field is required to meet its specifications everywhere in the far field region.

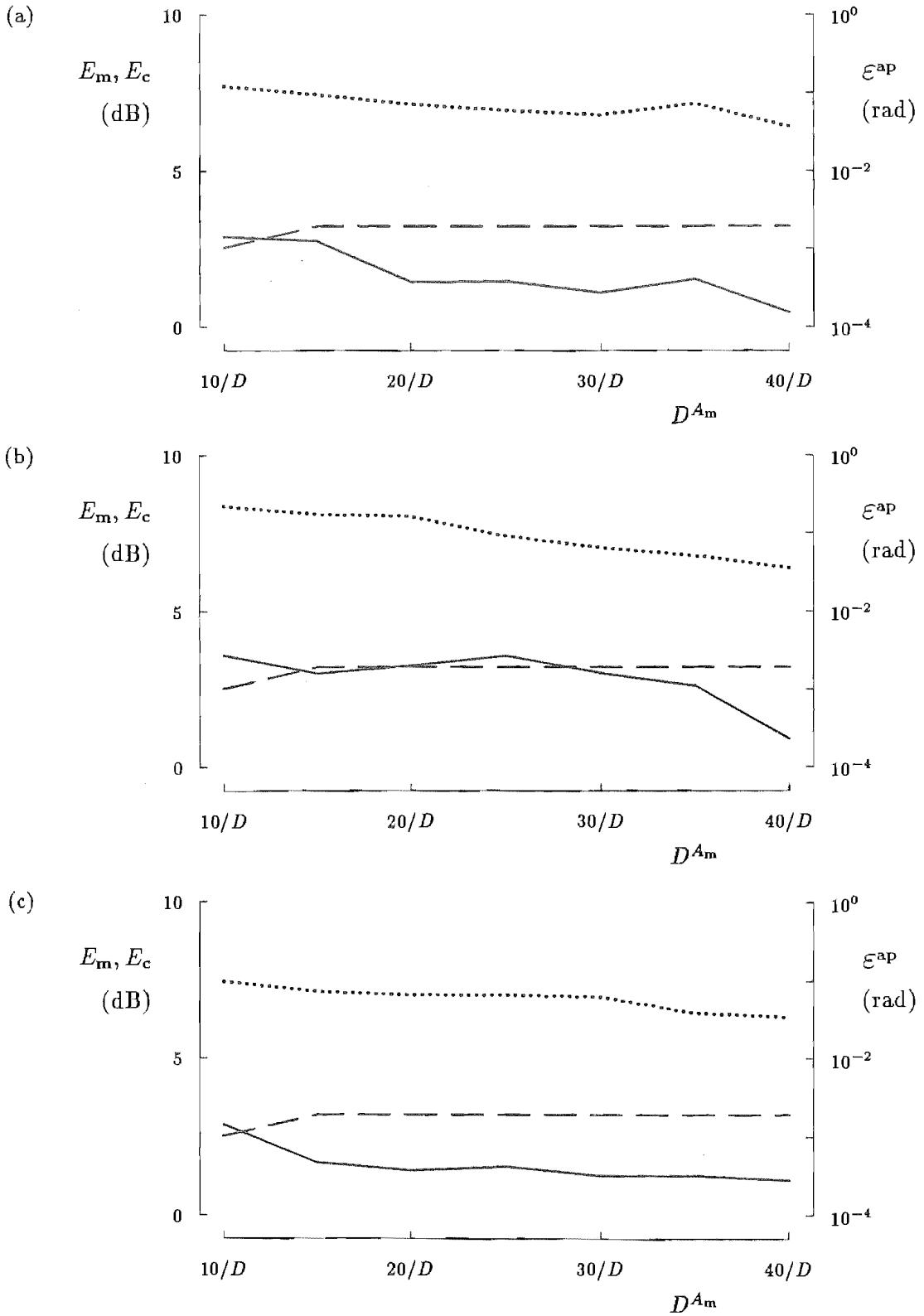


Figure 4.30 Errors for different approaches to dealing with truncated far field data: (a) composite algorithm applied directly to the truncated data; (b) composite algorithm applied to extrapolated data generated by the extrapolating smoothing algorithm; (c) extrapolating composite algorithm applied to the truncated data. Each graph shows values of $\mathcal{E}^{ap}$ (dotted curve), E_m (dashed curve) and E_c (solid curve) for each run. The particular model invoked for each run is the basic model but with one of 7 values of D^{A_m} in the range $10/D$ to $40/D$.

4.7.3.2 Extrapolating the far field data

In the same way that noisy data can be smoothed (Sec. 4.7.1), truncated data can be extrapolated. The problem of extrapolating $A_{\rm m}(u,v)$ is equivalent to the extrapolation problem posed in Section 3.5.6, because the inverse Fourier transform of $|F_{\rm a}(u,v)|^2$ is known to be compact, as noted in Section 4.7.1. Therefore, the extrapolation algorithm defined in Section 3.5.6 can be applied here. However, it is also known that $|F_{\rm a}(u,v)|^2$ is real and non-negative. One iteration of what is here called the *extrapolating smoothing algorithm* is defined by (cf. (3.89) and (4.49))

$$\begin{aligned}
H_{i_{\rm pre}}(u,v) &= \mathrm{FT}\{h_{i_{\rm pre}}(x,y)\} \\
H'_{i_{\rm pre}}(u,v) &= \begin{cases} [A_{\rm m}(u,v)]^2 & \text{if } (u,v) \in S^{A_{\rm m}} \\ H_{i_{\rm pre}}(u,v) & \text{if } (u,v) \notin S^{A_{\rm m}} \text{ and } H_{i_{\rm pre}}(u,v) \geq 0 \\ 0 & \text{if } (u,v) \notin S^{A_{\rm m}} \text{ and } H_{i_{\rm pre}}(u,v) < 0 \end{cases} \\
h'_{i_{\rm pre}}(x,y) &= \mathrm{IFT}\left\{H'_{i_{\rm pre}}(u,v)\right\} \\
h_{i_{\rm pre}+1}(x,y) &= \begin{cases} h'_{i_{\rm pre}}(x,y) & \text{for } (x,y) \in S^{\rm auto} \\ 0 & \text{elsewhere} \end{cases}
\end{aligned} \tag{4.53}$$

where $S^{\rm auto}$ is the autocorrelation support defined in Section 4.7.1. The algorithm is started by setting $h_1(x,y) = ff_{\rm m}(x,y)$, which is defined in (4.48). If the algorithm is run for $I_{\rm pre}$ iterations, $[H'_{I_{\rm pre}}(u,v)]^{1/2}$ is taken to be an extrapolated version of $A_{\rm m}(u,v)$. Note that $[H'_{I_{\rm pre}}(u,v)]^{1/2} = A_{\rm m}(u,v)$ within $S^{A_{\rm m}}$, implying that none of the measured samples are altered during the running of the algorithm.

Figure 4.30(b) depicts the errors for the composite algorithm applied to data produced by the extrapolating smoothing algorithm. For each run, the extrapolating algorithm was applied, for 15 iterations, to the truncated far field data $A_{\rm m}(u,v)$. These data were then replaced by $[H'_{15}]^{1/2}$, which was operated upon by the composite algorithm.

4.7.3.3 The extrapolating composite algorithm

An alternative way of extrapolating the far field data is to incorporate the extrapolation procedure into the composite algorithm. This is equivalent to solving the problem of finding a copolar aperture field distribution $f_{\rm e}(x,y)$ which is zero outside $S^{\rm aper}$ and whose far field amplitude pattern $|F_{\rm e}(x,y)|$ is almost equal to $A_{\rm m}(u,v)$ within $S^{A_{\rm m}}$. However, this problem does not have a unique solution [Byrne and Fiddy, 1987]. To reduce the number of possible solutions, an extra constraint is applied in the form of a threshold $\Gamma^{\rm f}_{\rm thres}$. The requirement is that $|F_{\rm e}(u,v)| \leq \Gamma^{\rm f}_{\rm thres}$ for all points (u,v) lying outside $S^{A_{\rm m}}$. Recall from Section 4.4.5 that the composite algorithm incorporates the CC and HIO algorithms, which are defined in Sections 4.4.2 and 4.4.3. Both of these latter algorithms are defined by (4.36), (4.38) and (4.31). The *extrapolating composite algorithm* is an extended form of the composite algorithm (Sec. 4.4.5), in which appropriate changes are made to these three equations, as discussed in the next paragraph.

A single iteration in the initial group of 400 iterations of the extrapolating CC algorithm is defined by (4.36), but with its second equation replaced by

$$G_i'(u,v) = \begin{cases} A_{\mathrm{m}}(u,v)e^{j\,\mathrm{phase}\{G_i(u,v)\}}e^{j\left|\mathrm{phase}\{G_{i-1}'(u,v)\}-\mathrm{phase}\{G_i(u,v)\}\right|} \\ \qquad\qquad \text{if } (u,v) \in S^{A_{\mathrm{m}}} \\ G_i(u,v) \\ \qquad\qquad \text{if } (u,v) \notin S^{A_{\mathrm{m}}} \text{ and } |GS_i(u,v)| \leq \Gamma^{\mathrm{f}}_{\mathrm{thres}} \\ [2\,\Gamma^{\mathrm{f}}_{\mathrm{thres}} - |G_i(u,v)|]e^{j\,\mathrm{phase}\{G_i(u,v)\}} \\ \qquad\qquad \text{if } (u,v) \notin S^{A_{\mathrm{m}}} \text{ and } \Gamma^{\mathrm{f}}_{\mathrm{thres}} < |GS_i(u,v)| < 2\,\Gamma^{\mathrm{f}}_{\mathrm{thres}} \\ 0 \\ \qquad\qquad \text{if } (u,v) \notin S^{A_{\mathrm{m}}} \text{ and } |GS_i(u,v)| \geq 2\,\Gamma^{\mathrm{f}}_{\mathrm{thres}} \end{cases} \tag{4.54}$$

By comparing (4.54) with the second equation of (4.45), it can be seen that the way in which the threshold constraint is applied, in the region of the far field plane outside $S^{A_{\mathrm{m}}}$, is the same as the way in which the threshold constraint is applied in the reflection algorithm, which is discussed in Section 4.4.6.3. Similarly, a single iteration in the initial group of 400 iterations of the extrapolating HIO algorithm is defined by (4.38) but with its second equation replaced by

$$G_i'(u,v) = \begin{cases} A_{\mathrm{m}}(u,v)e^{j\,\mathrm{phase}\{G_i(u,v)\}} \\ \qquad\qquad \text{if } (u,v) \in S^{A_{\mathrm{m}}} \\ G_i(u,v) \\ \qquad\qquad \text{if } (u,v) \notin S^{A_{\mathrm{m}}} \text{ and } |GS_i(u,v)| \leq \Gamma^{\mathrm{f}}_{\mathrm{thres}} \\ [2\,\Gamma^{\mathrm{f}}_{\mathrm{thres}} - |G_i(u,v)|]e^{j\,\mathrm{phase}\{G_i(u,v)\}} \\ \qquad\qquad \text{if } (u,v) \notin S^{A_{\mathrm{m}}} \text{ and } \Gamma^{\mathrm{f}}_{\mathrm{thres}} < |GS_i(u,v)| < 2\,\Gamma^{\mathrm{f}}_{\mathrm{thres}} \\ 0 \\ \qquad\qquad \text{if } (u,v) \notin S^{A_{\mathrm{m}}} \text{ and } |GS_i(u,v)| \geq 2\,\Gamma^{\mathrm{f}}_{\mathrm{thres}} \end{cases} \tag{4.55}$$

A single iteration in the final 100 iterations of either the extrapolating CC or HIO algorithm is defined by (4.31) but with its second equation also replaced by (4.55). When $S^{A_{\mathrm{m}}}$ encompasses the whole of the far field sampling grid, the extrapolating CC and HIO algorithms reduce respectively to the CC and HIO algorithms.

Results of running the extrapolating composite algorithm on many sets of truncated data, each defined by a different value of $D^{A_{\mathrm{m}}}$, are presented in Figure 4.30(c). The value of $\Gamma^{\mathrm{f}}_{\mathrm{thres}}$ was set to $0.0032A_{\mathrm{m}}(0,0)$ for all runs.

4.7.3.4 Comparison of approaches to dealing with truncated data

The three different approaches to utilizing truncated far field data, described in Sections 4.7.3.1 through 4.7.3.3, can be compared by inspecting Figure 4.30. The trends exhibited by $\mathcal{E}^{\mathrm{ap}}$ and E_{c} indicate that each of the approaches performs worse the smaller the value of $D^{A_{\mathrm{m}}}$. This is to be expected because the quantity of measured data available for generating an estimate $f_{\mathrm{e}}(x,y)$ of the image-form of $f_{\mathrm{a}}(x,y)$ decreases with $D^{A_{\mathrm{m}}}$.

From the values of $\mathcal{E}^{\mathrm{ap}}$ and E_{c} shown in Figure 4.30, the extrapolating smoothing algorithm is seen to perform worse, for most of the values of $D^{A_{\mathrm{m}}}$ which were invoked,

than either the extrapolating composite algorithm or direct application of the composite algorithm. The extrapolating composite algorithm approach performs better than direct application of the composite algorithm for some values of D^{A_m}, but performs worse for other values of D^{A_m}. Because the extrapolating composite algorithm produces more consistent results, I consider it to be the best overall approach as regards the particular models invoked to generate Figure 4.30.

4.8 ASSESSMENT OF COMPOSITE ALGORITHM

In the following sections, the composite algorithm, which is defined in Section 4.4.5, is evaluated by applying it to data generated from variety of particular models. Recall from Section 4.2 that any particular model is defined by a design and a set of parameters. When the value of an individual parameter is not its default value, it characterizes an individual imperfection, such as measurement noise or the aperture phase deviation due to a displaced panel. In order to assess the effect of an individual model parameter on the algorithm's convergence, the algorithm is applied to several different models whose definitions differ by only the value of that parameter. To illustrate the results, the values of $\mathcal{E}^{ap}$, E_m and E_c for each run of the composite algorithm is plotted against the value of the model parameter being considered.

The particular models utilized in Section 4.8.1 are relatively simple ones, so that the effect of one imperfection is isolated from the effect of related imperfections. The particular models utilized in Section 4.8.2 are variations of the basic model. The results presented in Section 4.8.2 show how the composite algorithm's convergence is affected by an individual imperfection in the presence of other imperfections. Section 4.8.3 presents results for a relatively comprehensive model which incorporates many simultaneously displaced panels. Included in these sections are studies of almost all of the parameters, introduced in Section 4.2, which define a particular model. Section 4.8.4 summarizes the results presented in Section 4.8.1 to 4.8.3.

4.8.1 Relatively simple computer models

Figures 4.31 to 4.36 indicate the degree of convergence of the composite algorithm when it is applied to data generated from relatively simple models. Each of these models incorporate design 1 and two imperfections. One imperfection is always an aperture phase deviation due to either defocus or panel displacement. The other imperfection is either an aperture amplitude deviation or a measurement inaccuracy.

The particular models invoked for Figures 4.31 to 4.36 are defined in Table 4.2. Note that, in each of the figures, the models invoked for subfigure (a) suffer from various amounts of defocus, the models invoked for subfigure (b) suffer from various amounts of panel displacement, while the models invoked for subfigure (c) suffer from displaced panels of different sizes. For those models in which there is no measurement inaccuracy, the envelope offset Γ_{off} is set to 0.002 so that the envelope errors E_m and E_c can be compared to those corresponding to the basic model.

Figures 4.31 and 4.32 depict results for the composite algorithm applied to models having a quadratic aperture amplitude deviation, characterized by $\tau_{quad} = 0.01$ and $\tau_{quad} = 0.1$ respectively. Note that, even when there is no aperture phase deviation $E_m = 0.2$ dB when $\tau_{quad} = 0.01$. As intimated in Section 4.3, the best possible value of E_c is therefore also 0.2 dB. From Figure 4.31 it can be seen that the composite algorithm converges to $E_c = 0.2$ dB for all but 3 of the runs represented in Figure 4.31.

A similar situation is illustrated by Figure 4.32. When there are no aperture phase deviations, the larger value of τ_{quad} implies a best possible value of E_c of 3.4 dB. In all but 4 of the runs represented in Figure 4.32, $E_c \leq 3.6$ dB. Note that, as expected and as indicated by the values of $\mathcal{E}^{ap}$ in Figures 4.31 and 4.32, the composite algorithm always generates a more accurate estimate of phase$\{f_a(x,y)\}$ when $\tau_{\text{quad}} = 0.01$ than when $\tau_{\text{quad}} = 0.1$.

Figures 4.33 and 4.34 depict results for the composite algorithm applied to data generated from models having a noise level, in the aperture field amplitude distribution, characterized by $\tau_{\text{ran}} = 0.005$ and $\tau_{\text{ran}} = 0.05$ respectively. Note that Figure 4.33 can be compared to Figure 4.31 because the rms value of $(|f_a(x,y)| - |f_d(x,y)|)$ when $\tau_{\text{ran}} = 0.005$ is approximately equal to that when $\tau_{\text{quad}} = 0.01$. For similar reasons, Figure 4.34 is comparable with Figure 4.32. For the smaller value of τ_{ran}, all but one of the runs converged to $E_c = 0.0$ dB. However, the convergence properties of the composite algorithm appear to be erratic when $\tau_{\text{ran}} = 0.05$.

Figures 4.35 and 4.36 show results for models in which the measurement noise is characterized by $\Gamma_{\text{ran}} = -60$ dB and $\Gamma_{\text{ran}} = -50$ dB respectively. For reasons discussed in Section 4.3, Γ_{off} is set to 0.002 and 0.006 for the models in the two respective figures. Notice that the value of $\mathcal{E}^{ap}$ to which the composite algorithm converges is relatively insensitive to the amount of aperture phase deviation, but is dependent on the level of far field measurement noise. Because Γ_{off} is introduced to offset the effects of Γ_{ran} on the envelope errors (Sec. 4.3), one might expect the algorithm always to converge to $E_c = 0$ dB. However, this is the case for only 20 of the 42 runs to which Figures 4.35 and 4.36 relate. This implies that, for approximately one half of the runs, the maximum value of $(|F_c(u,v)| - |F_d(u,v)|)/|F_d(0,0)|$ is greater than $\Gamma_{\text{off}} = 2\,\Gamma_{\text{ran}}$. All but two of these runs converged with $E_c < 1.0$ dB, which is, of course, satisfactory provided that the design safety margin E_{sm}, introduced in Section 4.3, is at least 1.0 dB.

Graph	Definitions of models
Figure 4.31(a)	$\tau_{\text{quad}} = 0.001$, design 1, $\Gamma_{\text{off}} = 0.002$, $\psi_{\text{quad}} = \{0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 32\}$ rad.
Figure 4.31(b)	$\tau_{\text{quad}} = 0.001$, design 1, $\Gamma_{\text{off}} = 0.002$, $\Omega_{\text{pan}} = 0.019$, $\psi_{\text{pan}} = \{0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2\}$ rad.
Figure 4.31(c)	$\tau_{\text{quad}} = 0.001$, design 1, $\Gamma_{\text{off}} = 0.002$, $\psi_{\text{pan}} = 0.6$ rad., $\Omega_{\text{pan}} = \{0.004, 0.008, 0.019, 0.036, 0.120, 0.251\}$
Figure 4.32(a)	$\tau_{\text{quad}} = 0.01$, design 1, $\Gamma_{\text{off}} = 0.002$, $\psi_{\text{quad}} = \{0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 32\}$ rad.
Figure 4.32(b)	$\tau_{\text{quad}} = 0.01$, design 1, $\Gamma_{\text{off}} = 0.002$, $\Omega_{\text{pan}} = 0.019$, $\psi_{\text{pan}} = \{0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2\}$ rad.
Figure 4.32(c)	$\tau_{\text{quad}} = 0.01$, design 1, $\Gamma_{\text{off}} = 0.002$, $\psi_{\text{pan}} = 0.6$ rad., $\Omega_{\text{pan}} = \{0.004, 0.008, 0.019, 0.036, 0.120, 0.251\}$
Figure 4.33(a)	$\tau_{\text{ran}} = 0.005$, design 1, $\Gamma_{\text{off}} = 0.002$, $\psi_{\text{quad}} = \{0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 32\}$ rad.
Figure 4.33(b)	$\tau_{\text{ran}} = 0.005$, design 1, $\Gamma_{\text{off}} = 0.002$, $\Omega_{\text{pan}} = 0.019$, $\psi_{\text{pan}} = \{0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2\}$ rad.
Figure 4.33(c)	$\tau_{\text{ran}} = 0.005$, design 1, $\Gamma_{\text{off}} = 0.002$, $\psi_{\text{pan}} = 0.6$ rad., $\Omega_{\text{pan}} = \{0.004, 0.008, 0.019, 0.036, 0.120, 0.251\}$
Figure 4.34(a)	$\tau_{\text{ran}} = 0.05$, design 1, $\Gamma_{\text{off}} = 0.002$, $\psi_{\text{quad}} = \{0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 32\}$ rad.
Figure 4.34(b)	$\tau_{\text{ran}} = 0.05$, design 1, $\Gamma_{\text{off}} = 0.002$, $\Omega_{\text{pan}} = 0.019$, $\psi_{\text{pan}} = \{0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2\}$ rad.
Figure 4.34(c)	$\tau_{\text{ran}} = 0.05$, design 1, $\Gamma_{\text{off}} = 0.002$, $\psi_{\text{pan}} = 0.6$ rad., $\Omega_{\text{pan}} = \{0.004, 0.008, 0.019, 0.036, 0.120, 0.251\}$
Figure 4.35(a)	$\Gamma_{\text{ran}} = -60$ dB, design 1, $\Gamma_{\text{off}} = 0.002$, $\psi_{\text{quad}} = \{0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 32\}$ rad.
Figure 4.35(b)	$\Gamma_{\text{ran}} = -60$ dB, design 1, $\Gamma_{\text{off}} = 0.002$, $\Omega_{\text{pan}} = 0.019$, $\psi_{\text{pan}} = \{0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2\}$ rad.
Figure 4.35(c)	$\Gamma_{\text{ran}} = -60$ dB, design 1, $\Gamma_{\text{off}} = 0.002$, $\psi_{\text{pan}} = 0.6$ rad., $\Omega_{\text{pan}} = \{0.004, 0.008, 0.019, 0.036, 0.120, 0.251\}$
Figure 4.36(a)	$\Gamma_{\text{ran}} = -50$ dB, design 1, $\Gamma_{\text{off}} = 0.006$, $\psi_{\text{quad}} = \{0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 32\}$ rad.
Figure 4.36(b)	$\Gamma_{\text{ran}} = -50$ dB, design 1, $\Gamma_{\text{off}} = 0.006$, $\Omega_{\text{pan}} = 0.019$, $\psi_{\text{pan}} = \{0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2\}$ rad.
Figure 4.36(c)	$\Gamma_{\text{ran}} = -50$ dB, design 1, $\Gamma_{\text{off}} = 0.006$, $\psi_{\text{pan}} = 0.6$ rad., $\Omega_{\text{pan}} = \{0.004, 0.008, 0.019, 0.036, 0.120, 0.251\}$

Table 4.2 Definition of particular models invoked to generate the results depicted in Figures 4.31 to 4.36. Where a set of values is listed within a pair of braces, each value pertains to a different particular model. Therefore, each row in the table defines seven models.

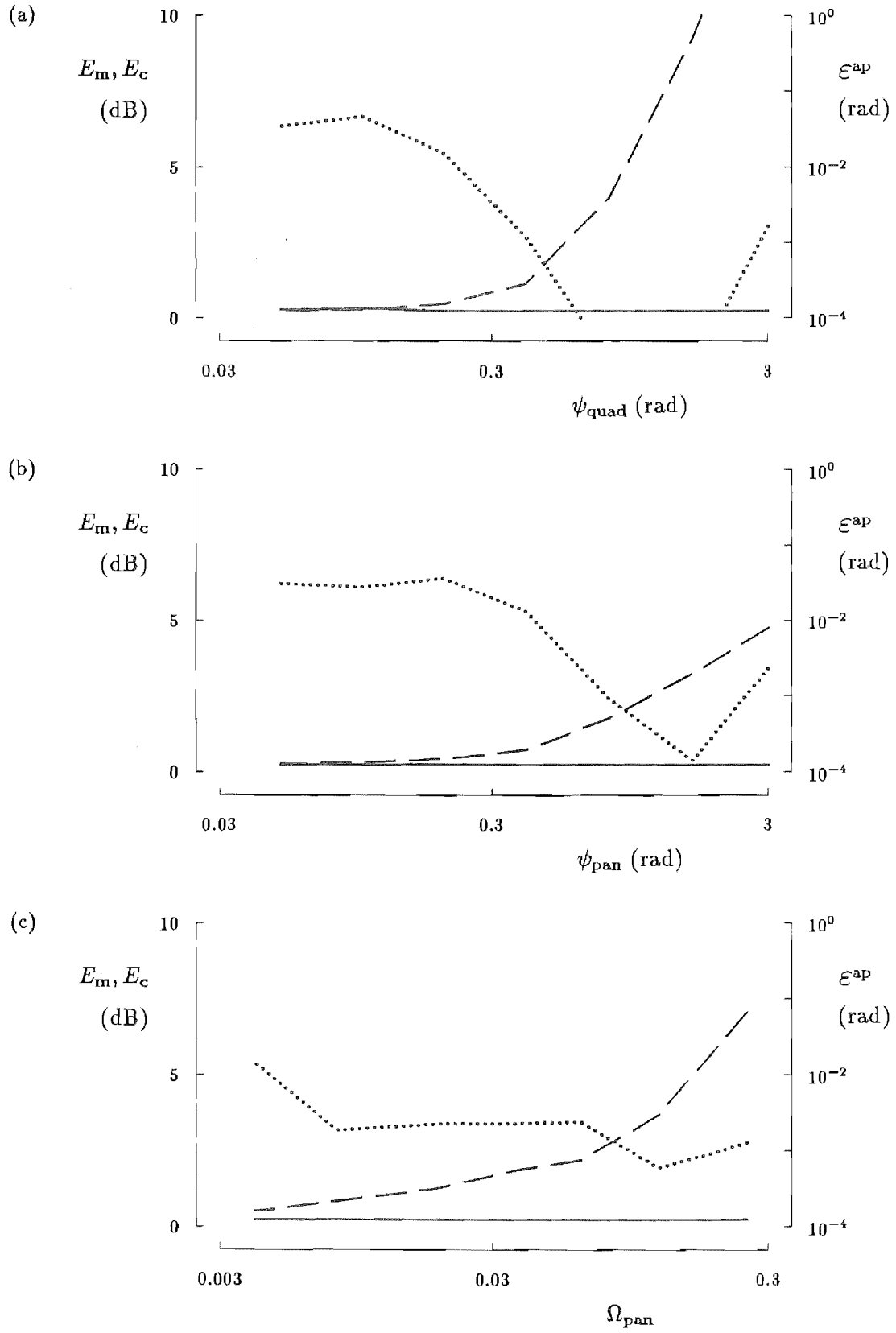


Figure 4.31 Effect of aperture phase deviations on convergence of composite algorithm when $\tau_{quad} = 0.01$. The particular models involved are listed in Table 4.2. Values of $\mathcal{E}^{ap}$ (dotted curve), E_m (dashed curve) and E_c (solid curve) are graphed for each run.

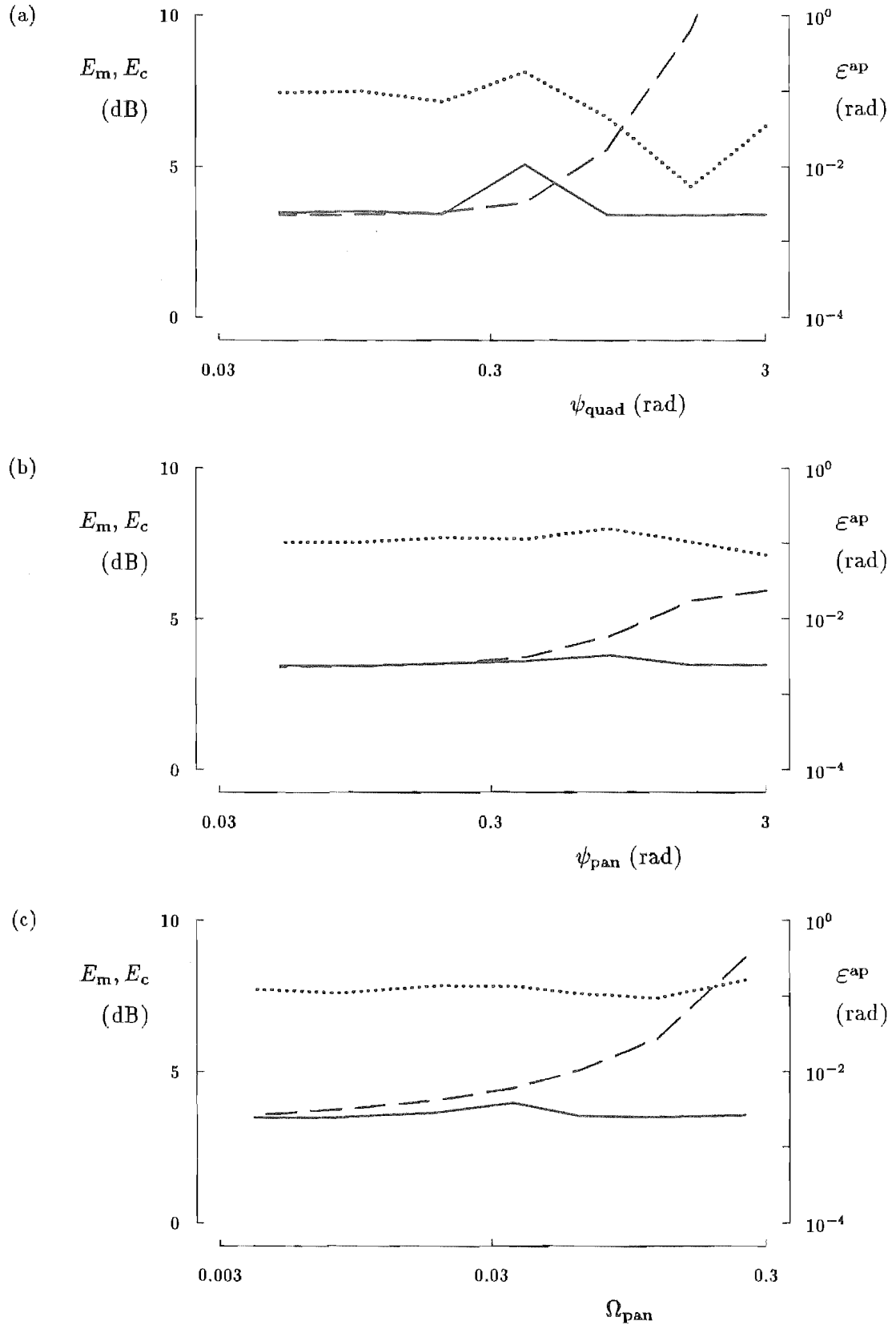


Figure 4.32 Effect of aperture phase deviations on convergence of composite algorithm when $\tau_{quad} = 0.1$. The particular models involved are listed in Table 4.2. Values of $\mathcal{E}^{ap}$ (dotted curve), E_m (dashed curve) and E_c (solid curve) are graphed for each run.

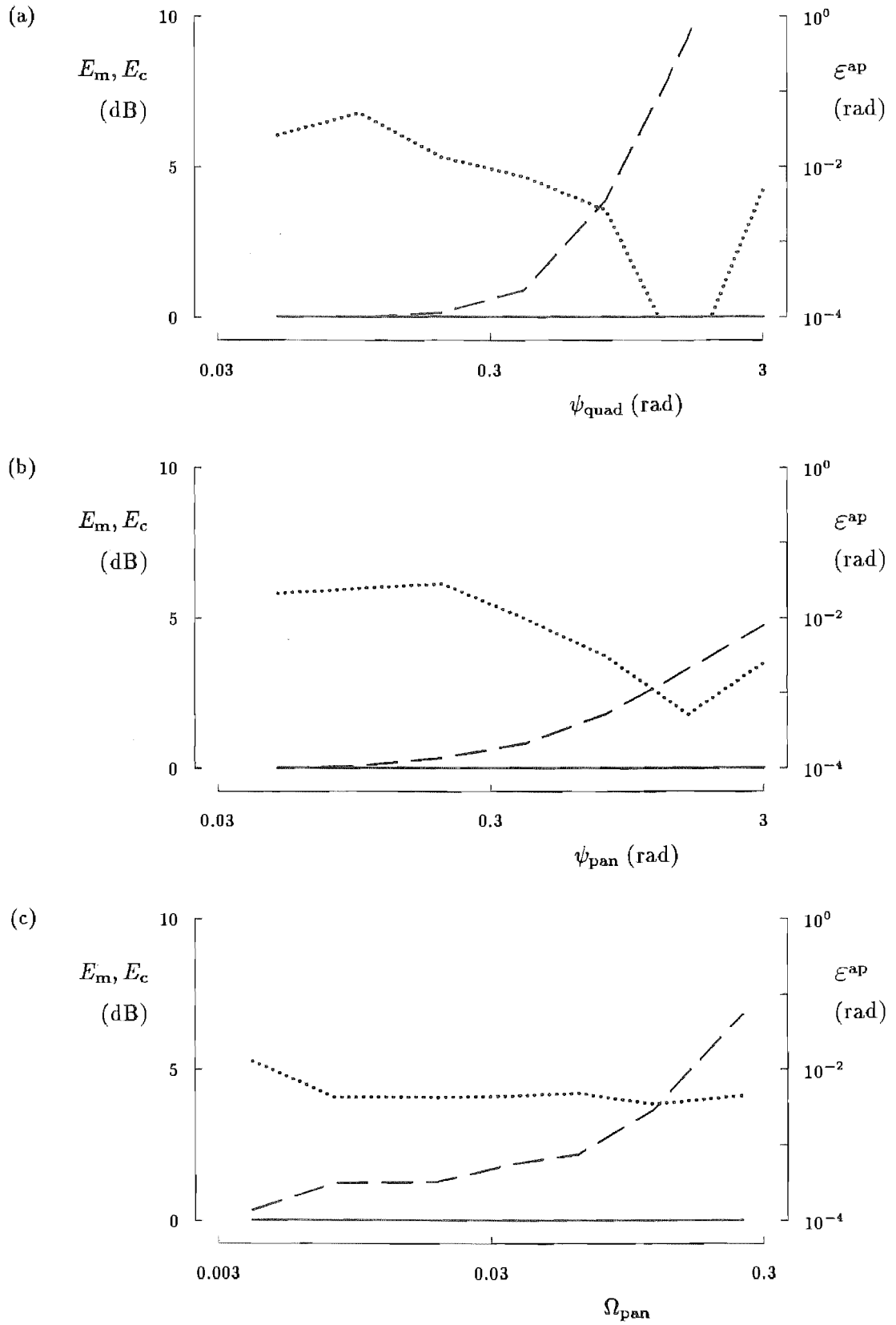


Figure 4.33 Effect of aperture phase deviations on convergence of composite algorithm when $\tau_{ran} = 0.005$. The particular models involved are listed in Table 4.2. Values of $\mathcal{E}^{ap}$ (dotted curve), E_m (dashed curve) and E_c (solid curve) are graphed for each run.

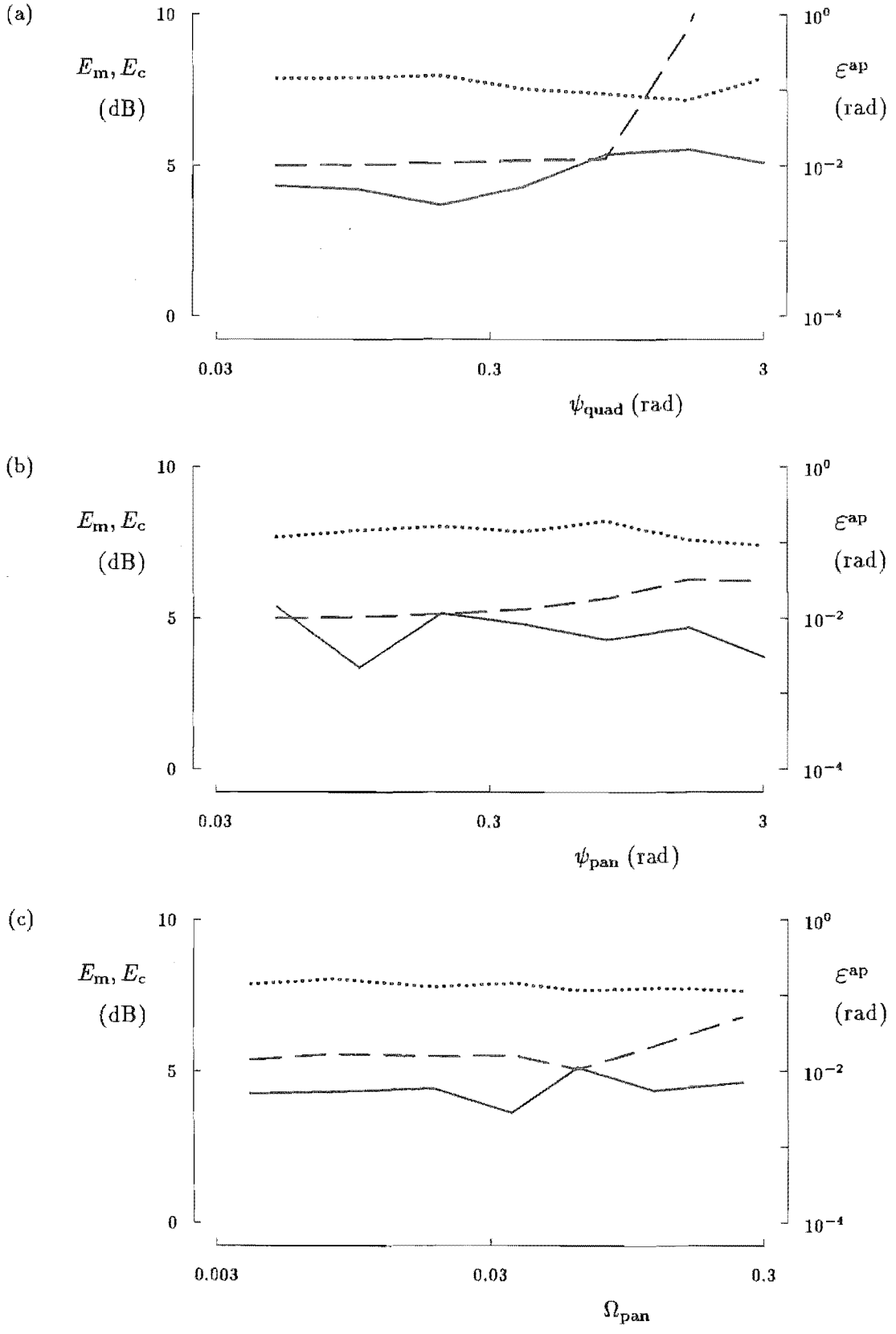


Figure 4.34 Effect of aperture phase deviations on convergence of composite algorithm when $r_{ran} = 0.05$. The particular models involved are listed in Table 4.2. Values of $\mathcal{E}^{ap}$ (dotted curve), E_m (dashed curve) and E_c (solid curve) are graphed for each run.

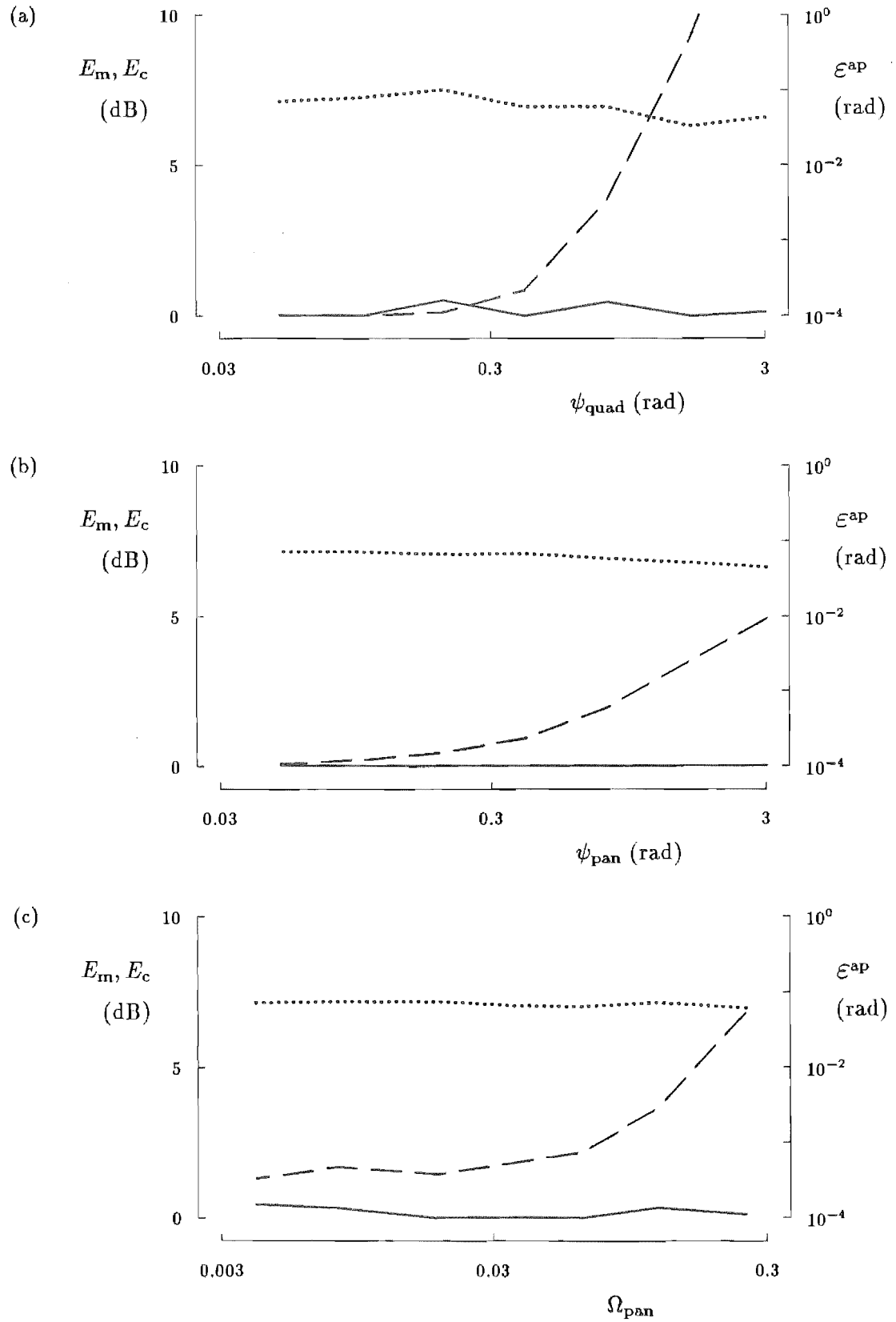


Figure 4.35 Effect of aperture phase deviations on convergence of composite algorithm when $\Gamma_{ran} = -60$ dB. The particular models involved are listed in Table 4.2. Values of $\mathcal{E}^{ap}$ (dotted curve), E_m (dashed curve) and E_c (solid curve) are graphed for each run.

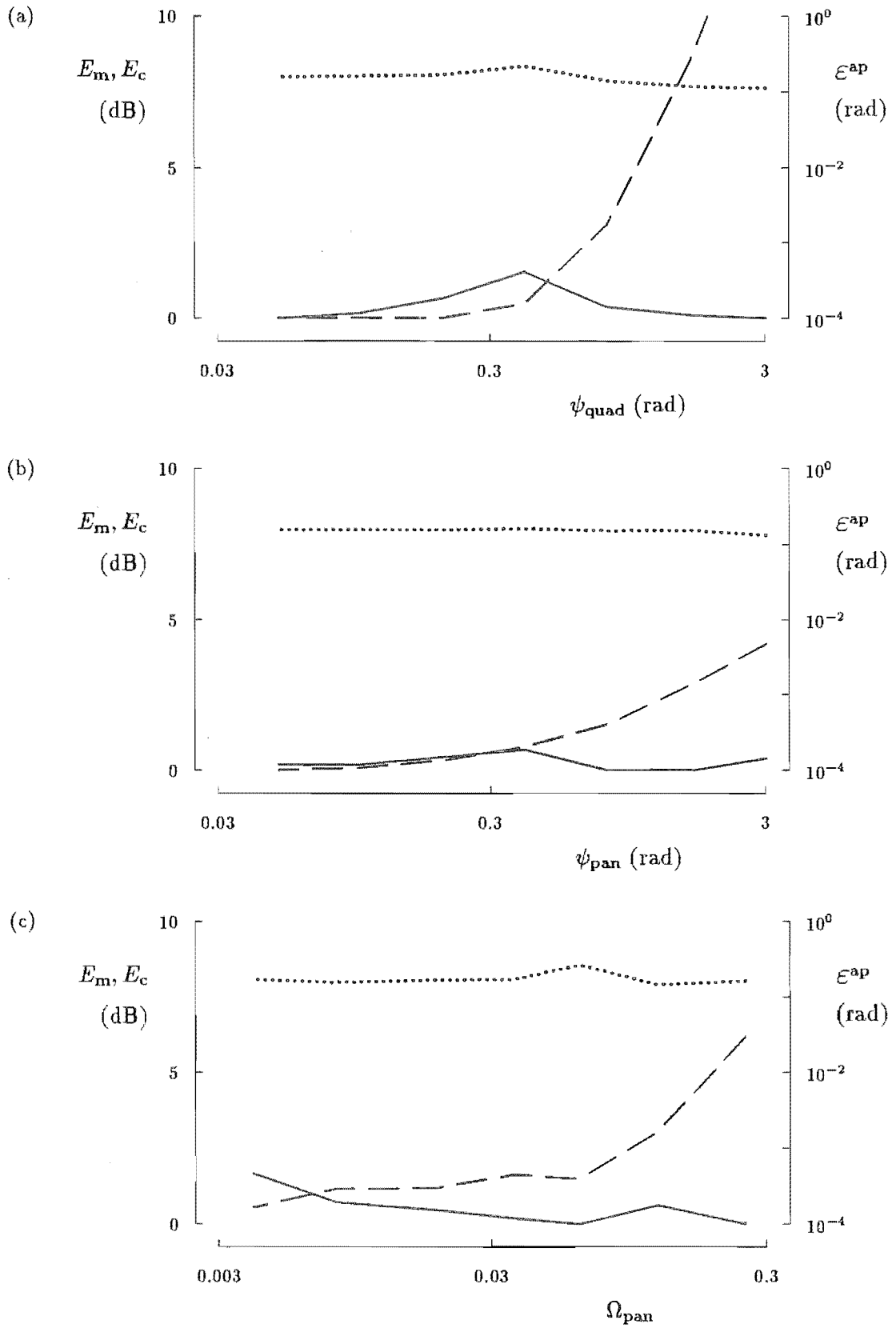


Figure 4.36 Effect of aperture phase deviations on convergence of composite algorithm when $\Gamma_{ran} = -50$ dB. The particular models involved are listed in Table 4.2. Values of $\mathcal{E}^{ap}$ (dotted curve), E_m (dashed curve) and E_c (solid curve) are graphed for each run.

4.8.2 Variations of the basic model

The simple models invoked to generate the results presented in the previous section are not very realistic. In practice, aperture phase deviations, aperture amplitude deviations and measurement inaccuracies all exist together. Therefore, the approach taken for the results presented in this section, is to invoke perturbations to the basic model, which is defined by (4.47). Note that the basic model incorporates aperture amplitude and phase deviations and measurement noise.

The effect of measurement noise Γ_{ran} on the convergence of the composite algorithm is depicted in Figure 4.37. The results were generated by applying the algorithm nine times to the basic model, but with Γ_{ran} set to a different value in the range -80 to -40 dB for each run. The value of Γ_{off} was set for all runs to be 3 dB greater than Γ_{ran}. Therefore, for larger values of Γ_{ran}, the levels of the design envelope A_{d} increase, thereby reducing the amount by which $A_{\text{m}}(u, v)$ exceeds A_{d}. The implication is that E_{m} decreases with increasing values of Γ_{ran}, as can be observed in Figure 4.37. Another trend apparent in Figure 4.37 is that $\mathcal{E}^{\text{ap}}$ is approximately proportional to Γ_{ran}: e.g. a ten fold increase in Γ_{ran} causes an approximately ten fold increase in $\mathcal{E}^{\text{ap}}$. Morris [1985] notes that, for radio measurements, the dominant error is likely to be additive receiver noise. Therefore, the phase accuracy of $f_{\text{e}}(x, y)$ generated by the composite algorithm is likely to be limited by the level of measurement noise. For all but one of the runs, E_{c} is less than E_{m} and is also less than 2 dB.

Figure 4.38 depicts results generated by the basic model with varying levels of calibration inaccuracy Γ_{cal}. Note that, when $\Gamma_{\text{cal}} \neq 1.0$, the particular models incorporate both calibration inaccuracy and measurement noise. From the way that calibration inaccuracy is simulated by (4.15), a value of Γ_{cal} greater than unity causes the sidelobe

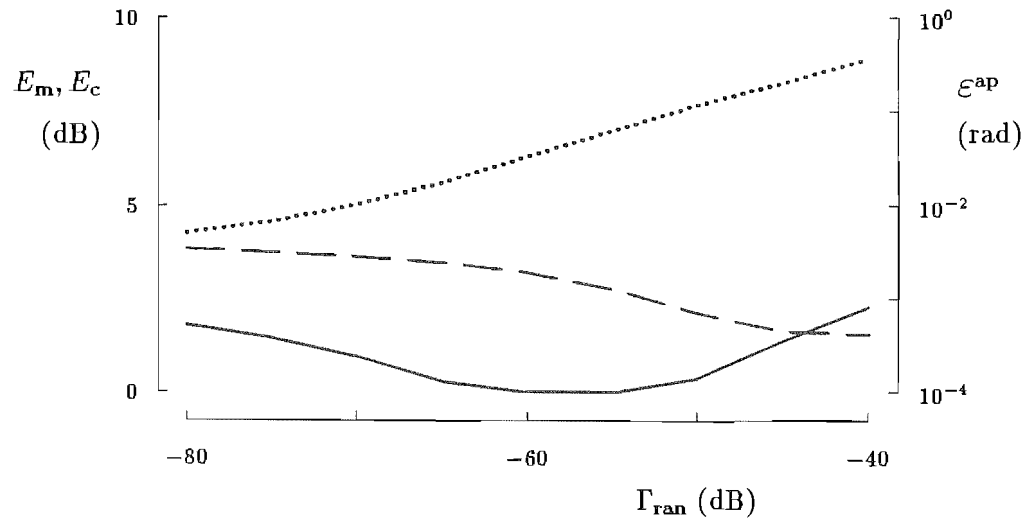


Figure 4.37 Effect of far field measurement noise Γ_{ran} on convergence of the composite algorithm. The models invoked to generate the data are described by the basic model but with $\Gamma_{\text{ran}} = -80, -75, -70, \ldots, -40$ dB respectively. The value of Γ_{off} in each case is double that of Γ_{ran} when neither is expressed in dB. Values of $\mathcal{E}^{\text{ap}}$ (dotted curve), E_{m} (dashed curve) and E_{c} (solid curve) are graphed for each run.

levels of $A_m(u,v)$, relative to the peak level, to be less than those of $|F_a(u,v)|$. When Γ_{cal} is less than unity, the sidelobe levels of $A_m(u,v)$ are lower than those of $|F_a(u,v)|$. This is why the measured envelope error E_m reduces as the value of Γ_{cal} increases. However, the value of E_c is determined by $|F_c(u,v)|$, which does not suffer from calibration inaccuracy. Thus, for these results, E_c should not be compared directly with E_m. As indicated by both E_c and $\mathcal{E}^{ap}$, the composite algorithm converges worse the further is the value of Γ_{cal} from unity. However, the convergence characteristics favour values of Γ_{cal} which are greater than unity. The results presented in Figure 4.38 therefore suggest that it is preferable to have a calibration inaccuracy which causes the sidelobes of $A_m(u,v)$ to be suppressed rather than enhanced.

The results depicted in Figure 4.39 indicate the effect on the composite algorithm's convergence of smoothly varying differences between the amplitudes of the actual and design copolar aperture distributions. From (4.9) and (4.13), a positive value of τ_{quad} implies that $|f_a(x,y)|$ is more tapered towards the aperture's edge than is $|f_d(x,y)|$. Various values of τ_{quad} were incorporated into the basic model to generate the results depicted in Figure 4.39. Figure 4.39 shows that E_m decreases as τ_{quad} increases. This is because the high sidelobe levels of $|F_a(u,v)|$, due to the aperture phase deviations in $f_a(x,y)$, are reduced by the increased tapering of $|f_a(x,y)|$. Figure 4.39 demonstrates that the composite algorithm converges to a higher value of $\mathcal{E}^{ap}$ the greater is τ_{quad}.

Figure 4.40 indicates how the convergence of the composite algorithm is affected by the level of complex noise τ_{ran} in the copolar aperture field distribution. As intimated in Section 4.2.2, the effect on $A_m(u,v)$ of the complex noise is to increase the sidelobe levels. This explains why it can be observed in Figure 4.40 that the value of E_m rises with increasing values of τ_{ran}. It is apparent from the curve for E_c in Figure 4.40 that, provided $\tau_{ran} < 0.02$, the composite algorithm converges as well as can be expected. However, the algorithm converges significantly worse when $\tau_{ran} \geq 0.05$.

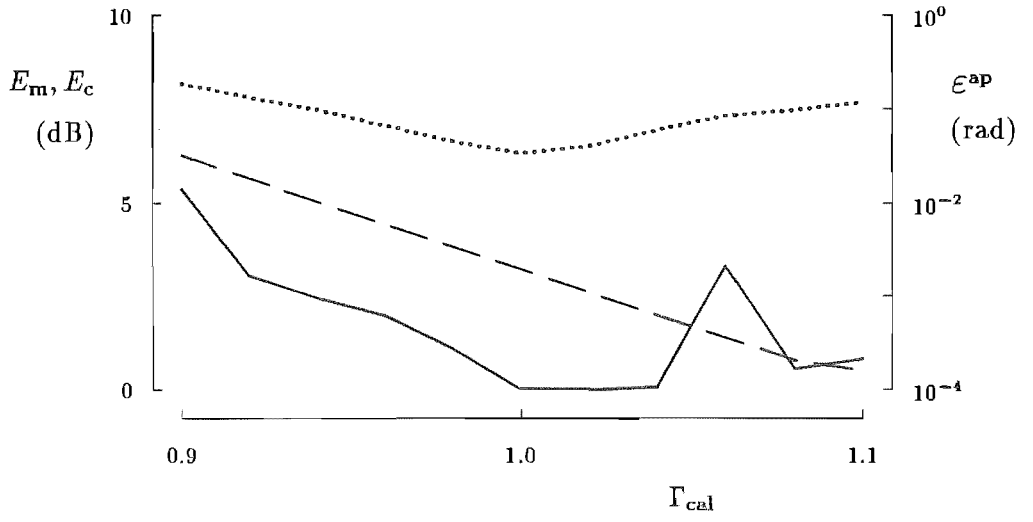


Figure 4.38 Effect of far field calibration inaccuracy Γ_{cal} on convergence of the composite algorithm. The models invoked to generate the data are described by the basic model supplemented by $\Gamma_{cal} = 0.9, 0.92, 0.94, \ldots, 1.1$ respectively. Values of $\mathcal{E}^{ap}$ (dotted curve), E_m (dashed curve) and E_c (solid curve) are graphed for each run.

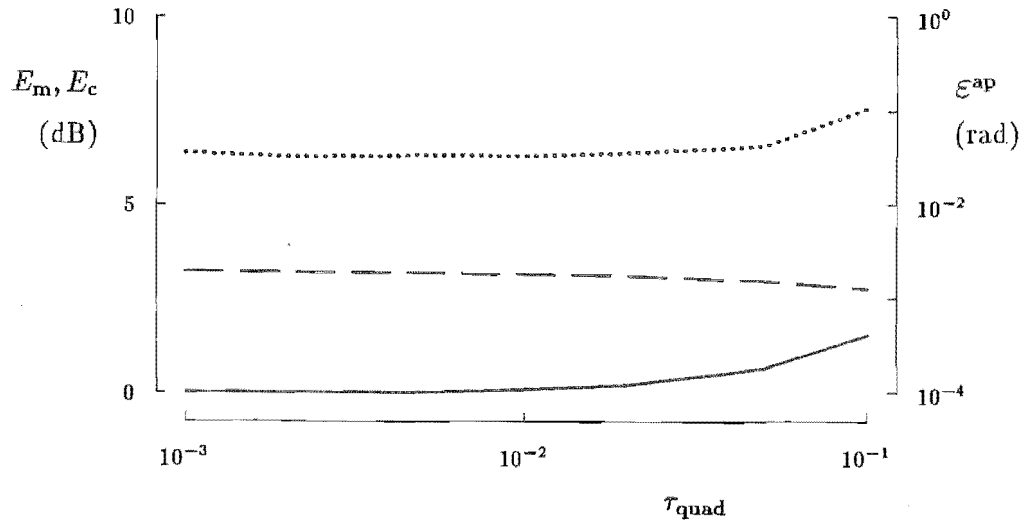


Figure 4.39 Effect of smooth aperture amplitude deviations, characterized by τ_{quad}, on convergence of the composite algorithm. The models invoked to generate the data are described by the basic model supplemented by $\tau_{quad} = 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1$ respectively. Values of $\mathcal{E}^{ap}$ (dotted curve), E_m (dashed curve) and E_c (solid curve) are graphed for each run.

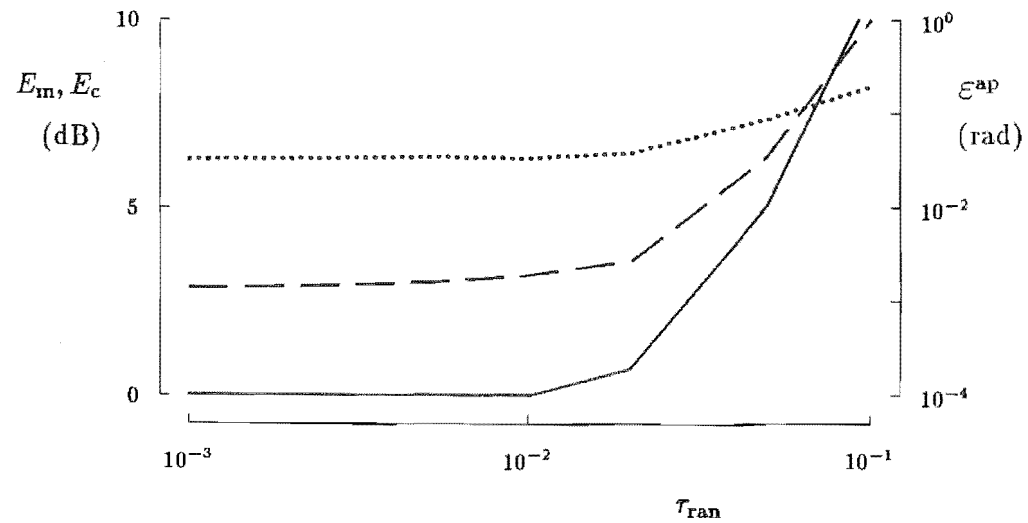


Figure 4.40 Effect of aperture field noise, characterized by τ_{ran}, on convergence of the composite algorithm. The models invoked to generate the data are described by the basic model but with $\tau_{ran} = 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1$ respectively. Values of $\mathcal{E}^{ap}$ (dotted curve), E_m (dashed curve) and E_c (solid curve) are graphed for each run.

4.8.3 Relatively comprehensive computer models

In this section, four examples are given which involve models more comprehensive than those invoked in the previous two sections. All four of the models invoked here incorporate the effect of many simultaneously displaced panels. Because this cannot be simulated by the model described in Section 4.2, (4.10) is replaced by

$$\Delta\psi(x,y) = \psi_{\text{quad}}\,\rho^2 + P_{\text{many}}(x,y) \tag{4.56}$$

where $P_{\text{many}}(x,y)$ simulates the aperture phase deviation due to the displaced panels. The geometry of the panels is taken to be that used by Bennett and Godwin [1977]. $P_{\text{many}}(u,v)$ is depicted in Figure 4.41. Most of the displaced panels produce a uniform phase deviation over the region corresponding to the projection of the panel onto the aperture plane. In these regions, $P_{\text{many}}(x,y)$ equals either 0.63 or -0.63 rad. Two of the panels are tilted so that they each produce a sloped phase distribution which is 0.63 rad. at one end, but -0.63 rad. at the other end, of the region in the aperture plane corresponding to the panel. There are also five circularly symmetric Gaussian distributions incorporated into $P_{\text{many}}(x,y)$ which each have an peak value of 0.2 rad. They simulate the effect of dents in the main reflector.

The equivalent of the basic model is defined by (cf. (4.47))

$$\begin{aligned} &\text{design 2;} && P_{\text{many}}(x,y), \\ &\psi_{\text{quad}} = 1.0 \text{ rad.;} && \tau_{\text{ran}} = 0.01; \\ &\Gamma_{\text{ran}} = 0.001 = -60 \text{ dB;} && \Gamma_{\text{off}} = 0.002 \end{aligned} \tag{4.57}$$

where the mention of $P_{\text{many}}(x,y)$ denotes that this particular model incorporates the effect of many displaced panels, as outlined in the previous paragraph. The results for the composite algorithm applied to this model are displayed in Figure 4.42(a). The aperture phase deviations generate a relatively large value of E_{m}. The algorithm converges to a value of $\mathcal{E}^{\text{ap}}$ which is commensurate with the level of measurement noise incorporated into this particular model. However the value of E_{c} is 0.3 dB, which is larger than might be expected, based on the value of $\mathcal{E}^{\text{ap}}$ and results already presented in this section which have similar values of $\mathcal{E}^{\text{ap}}$. However, this value of E_{c} represents a substantial improvement over the value of E_{m}.

The results presented in Figure 4.42(b) are generated by the application of the composite algorithm to a particular model described by (4.57) supplemented with $\tau_{\text{quad}} = 0.05$ and $\Gamma_{\text{cal}} = 1.02$. As for the results shown in Figure 4.42(a), the results in Figure 4.42(b) show that the value of E_{c} represents a considerable improvement over the value of E_{m}.

However, the composite algorithm does not always converge as well as in the previous two examples. Figure 4.42(c) depicts the results associated with the model described in (4.57), but with no quadratic aperture phase deviation. The model is therefore defined by

$$\begin{aligned} &\text{design 2,} && P_{\text{many}}(x,y), \\ &\tau_{\text{ran}} = 0.01, && \Gamma_{\text{ran}} = 0.001 = -60 \text{ dB}, \quad \Gamma_{\text{off}} = 0.002 \end{aligned} \tag{4.58}$$

The value of $\mathcal{E}^{\text{ap}}$, depicted in Figure 4.42(c), is higher than might be expected for the level of measurement noise incorporated into this model. This has the related effect that E_{c} is greater than E_{m}. One way of encouraging the composite algorithm to converge

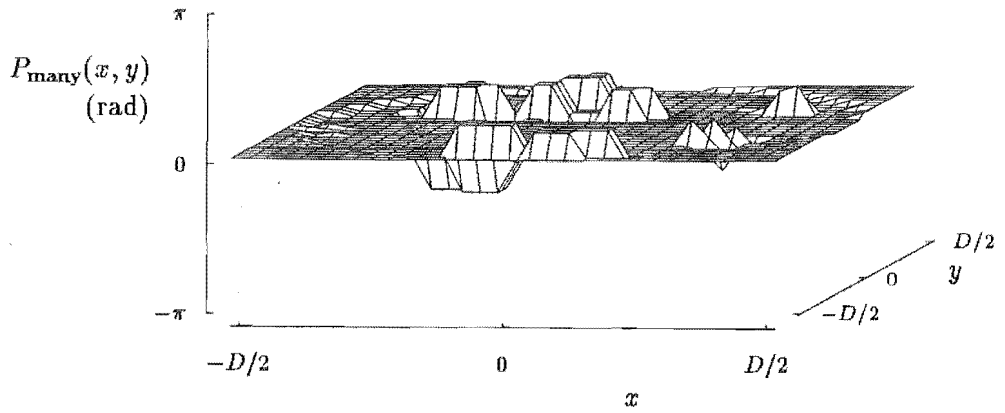


Figure 4.41 Aperture phase deviation $P_{\text{many}}(x, y)$ produced by many displaced panels. In order to discern the perimeter of the aperture, $P_{\text{many}}(x, y)$ has been set to 0.1 rad. for $(x, y) \notin S^{\text{aper}}$. The geometry of these panels is defined by Bennett and Godwin [1977].

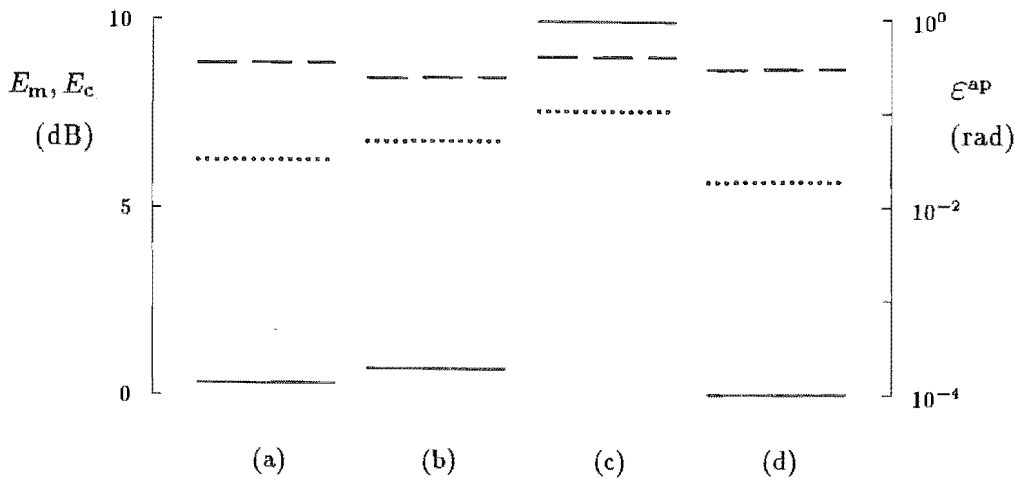


Figure 4.42 Results of the composite algorithm applied to relatively comprehensive models. The models involved are defined (a) by (4.57), (b) by (4.57) supplemented with $\tau_{\text{quad}} = 0.05$ and $\Gamma_{\text{cal}} = 1.02$, (c) (4.58), (d) (4.58) but with $\Gamma_{\text{ran}} = -70$ dB. Values of $\mathcal{E}^{\text{ap}}$ (dotted line), E_{m} (dashed line) and E_{c} (solid line) are indicated for each run.

better is to apply it to data which is more accurate. Accordingly, the model invoked for Figure 4.42(d) is defined by (4.58) but with $\Gamma_{\text{ran}} = -70$ dB instead of -60 dB. The value of Γ_{off} is kept at 0.03, so that the envelope errors in Figures 4.42(c) and (d) can be directly compared with each other. The results in Figures 4.42(d) show that, for this particular example, increasing the accuracy of the measured copolar far field amplitude pattern enables the composite algorithm to converge much better.

4.8.4 Summary of results

This section summarizes the results presented in Sections 4.8.1 to 4.8.3.

The values of E_c plotted in Figures 4.31 to 4.36 are encouraging because they indicate that the degree of convergence of the composite algorithm does not usually depend upon the amount and characteristics of the aperture phase deviation between $f_a(x, y)$ and $f_d(x, y)$. The plotted values of $\mathcal{E}^{ap}$ also indicate that the convergence properties are independent of the aperture phase deviations in the presence of measurement inaccuracies. In the ideal situation, however, for which there are no measurement inaccuracies, the composite algorithm generates a more faithful estimate of phase$\{f_a(x, y)\}$ the larger the aperture phase deviations are.

It is apparent from all of the results presented in Sections 4.8.1 to 4.8.3 that the convergence properties of the composite algorithm are dependent upon the accuracy of the data to which it is applied. The greater the far field measurement inaccuracy (characterized by Γ_{ran} and Γ_{cal}) the greater are the values of both $\mathcal{E}^{ap}$ and E_c. Similarly, the greater the aperture amplitude deviations (characterized by τ_{quad} and τ_{ran}) between $f_a(x, y)$ and $f_d(x, y)$, the greater is the value of $\mathcal{E}^{ap}$. Recall from Section 4.3 that E_c is affected not only by the accuracy, indicated by $\mathcal{E}^{ap}$, of phase$\{f_e(x, y)\}$, but also by the aperture amplitude deviations (see (4.30)). The values of E_c plotted in Figures 4.31 to 4.36 indicate that the former effect on E_c is often negligible compared to the latter effect. However, this is not always the case, especially for models having relatively large values of τ_{quad} and τ_{ran}.

Out of a total of 164 different simulations presented in Sections 4.8.1 to 4.8.3, the composite algorithm failed outright, but converged with a value of E_c exceeding E_m, for 25 of the simulations. Of these 25 failed simulations, 16 are associated with Figures 4.32, 4.34 or 4.36, for which $\tau_{\text{quad}} = 0.1$, $\tau_{\text{ran}} = 0.05$ and $\Gamma_{\text{ran}} = -50$ dB respectively. All of these parameters imply that the composite algorithm was applied to relatively inaccurate aperture or far field data.

This paragraph discusses only those runs of the composite algorithm applied to models for which $\tau_{\text{quad}} < 0.1$, $\tau_{\text{ran}} < 0.05$, $\Gamma_{\text{ran}} < -50$ dB and $|\Gamma_{\text{cal}} - 1.0| < 0.05$. This leaves 89 of the simulations presented in Section 4.8.1 to 4.8.3. In 50% of these simulations, the composite algorithm converged to $E_c = 0.0$ dB, which represents the best convergence possible. The algorithm converged to within $E_c = 0.5$ dB and 1 dB for, respectively, 89% and 94% of the simulations. All but one of the runs converged with $E_c < 2$ dB. With regard to the aperture phase error $\mathcal{E}^{ap}$, the algorithm converged to within 0.04, 0.06 and 0.08 rad. for 67%, 82% and 99% of the simulations. From the discussion following (4.29), these values of $\mathcal{E}^{ap}$ imply that, ignoring the phase deviations caused by scattering from the struts, the shape of the main reflector can be calculated to within tolerances of about $\lambda/300$, $\lambda/200$ and $\lambda/150$ respectively.

4.9 OTHER USES OF THE MODIFIED GERCHBERG-SAXTON ALGORITHM

The modified Gerchberg-Saxton algorithm can be utilized to generate estimates of more than just the copolar aperture field phase distribution. In the next two sections, two other applications of the Gerchberg-Saxton algorithm are investigated. Section 4.9.1 discusses the use of phase retrieval to determine the tilt angle of a linearly polarized feed so that feed can be adjusted to generate minimal depolarization. The accuracy of the amplitude of the copolar aperture field distribution generated by the composite algorithm is investigated in Section 4.9.2. Knowing the copolar aperture field amplitude distribution aids in diagnosing situations such as the feed radiating a more tapered field than it is designed to.

4.9.1 Estimation of depolarization

In many situations in which high gain microwave antennas are employed, depolarization at the centre of the far field radiation pattern is undesirable (e.g. see Sec. 2.4.2.1 on frequency reuse in satellite communications). As mentioned in Section 4.2.4, depolarization is not only due to the tilt of the antenna's feed, which is assumed here to be linearly polarized, but is also due to aperture phase deviations and the cross polar component of the field radiated by the feed. The modified Gerchberg-Saxton algorithm can be invoked to generate, from $A_{\mathrm{m}}(u,v)$ and $|f_{\mathrm{d}}(x,y)|$, an estimate phase$\{f_{\mathrm{e}}(x,y)\}$ of the aperture phase deviations. However, before the feed tilt can be estimated, further information is required. This section describes a technique [Gardenier *et al.*, 1988] for estimating the feed tilt from phase$\{f_{\mathrm{e}}(x,y)\}$, which is generated by the modified Gerchberg-Saxton algorithm, given the design cross polar aperture field aperture distribution $|f_{\mathrm{d}}^{\mathrm{x}}(x,y)|$ and the measured cross polar far field amplitude pattern $A_{\mathrm{m}}(u,v)$. This technique is illustrated with a computer simulated example.

The notation employed throughout this section is that developed in Section 4.2.4. To simplify the following explanation, all noise terms are initially neglected (i.e. τ_{ran} and Γ_{ran} are set to zero). It is required that, before the following algorithm is invoked, the modified Gerchberg-Saxton algorithm be run, to generate an estimate $f_{\mathrm{e}}(x,y)$ of $f_{\mathrm{a}}(x,y)$.

Substituting the second equation of (4.18) into the first equation of (4.19) yields

$$f_{\mathrm{a}}(x,y) = [1 + \phi_{\mathrm{tilt}}\, \tau_{\mathrm{xpol}} \sin(2\phi)]\, f_{\mathrm{nt}}(x,y) \tag{4.59}$$

Recall that, as defined in Section 4.2.4, $f_{\mathrm{nt}}(x,y)$ would have been the actual copolar aperture field distribution had ϕ_{tilt} equalled zero. Since $\phi_{\mathrm{tilt}}\, \tau_{\mathrm{xpol}} \sin(2\phi)$ is real and, for all values of ϕ, is always greater than -1, phase$\{f_{\mathrm{nt}}(x,y)\}$ = phase$\{f_{\mathrm{a}}(x,y)\}$ and is therefore approximated by phase$\{f_{\mathrm{e}}(x,y)\}$. It follows from the (4.18) that, assuming $\Delta a(x,y)$ is negligible, estimates $f_{\mathrm{ent}}(x,y)$ and $f_{\mathrm{ent}}^{\mathrm{x}}(x,y)$ of $f_{\mathrm{nt}}(x,y)$ and $f_{\mathrm{nt}}^{\mathrm{x}}(x,y)$, respectively, are given by

$$\begin{aligned} f_{\mathrm{ent}}(x,y) &= f_{\mathrm{d}}(x,y) e^{j\,\mathrm{phase}\{f_{\mathrm{e}}(x,y)\}} \\ f_{\mathrm{ent}}^{\mathrm{x}}(x,y) &= f_{\mathrm{d}}^{\mathrm{x}}(x,y) e^{j\,\mathrm{phase}\{f_{\mathrm{e}}(x,y)\}} \end{aligned} \tag{4.60}$$

The Fourier transform of the second equation of (4.19) is

$$F_{\mathrm{a}}^{\mathrm{x}}(u,v) = F_{\mathrm{nt}}^{\mathrm{x}}(u,v) - \phi_{\mathrm{tilt}}\, F_{\mathrm{nt}}(u,v) \tag{4.61}$$

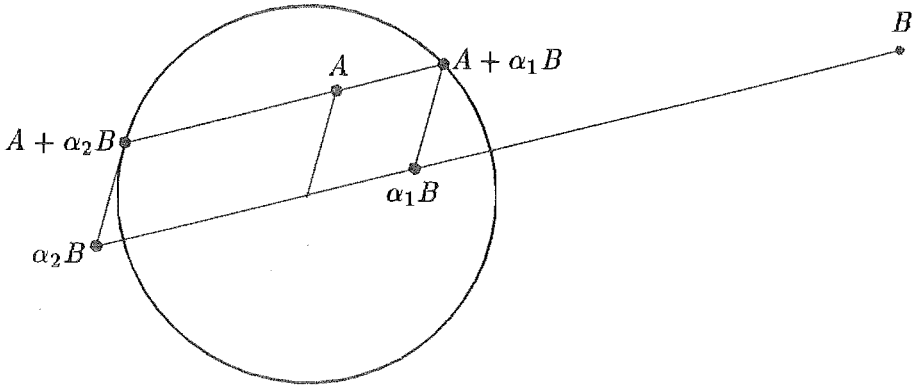


Figure 4.43 Complex plane showing the two solutions for α for the equation $|A + \alpha B| = |C|$, where A, B and $|C|$ are known. This equation is equivalent to (4.62) where A and B are complex numbers and $|C|$ and α are real values.

In words, this equation means that a tilted feed adds a (negative) proportion of the copolar radiation pattern into the cross polar radiation pattern. Whether or not the patterns reinforce or cancel, at a particular point in the u, v plane, depends on their relative phases at that point.

As indicated by (4.20), the amplitude of $F_{\mathrm{a}}^{\times}(u, v)$ is equal to $A_{\mathrm{m}}^{\times}(u, v)$, remembering that measurement noise has been ignored. Therefore, after replacing $F_{\mathrm{nt}}(u, v)$ and $F_{\mathrm{nt}}^{\times}(u, v)$ with their estimates, taking the amplitude of both sides of (4.61) yields

$$A_{\mathrm{m}}^{\times}(u, v) = |F_{\mathrm{ent}}^{\times}(u, v) - \phi_{\mathrm{tilt}} F_{\mathrm{ent}}(u, v)| \tag{4.62}$$

in which ϕ_{tilt} is the only unknown. Squaring both sides of (4.62) yields a quadratic function in ϕ_{tilt}. Therefore, at each far field sample point, (4.62) can be solved to yield two solutions for ϕ_{tilt}, which are illustrated geometrically in Figure 4.43. When several sample points are considered, one solution for each sample point is equal to ϕ_{tilt}, so these solutions must equal each other. The other solutions for ϕ_{tilt} are expected to vary from sample point to sample point, because the relative values of $F_{\mathrm{ent}}(u, v)$ and $F_{\mathrm{ent}}^{\times}(u, v)$ vary also. Therefore, by solving (4.62) to obtain a pair of solutions for ϕ_{tilt} at each of two or more sample points and taking the common solution from each pair, the correct solution for ϕ_{tilt} is obtained.

However, when measurement noise is appreciable, the expression for $A_{\mathrm{m}}^{\times}(u, v)$ in (4.62) holds only approximately. Therefore, solving (4.62) at several far field sample points generates one solution from each sample point which is an approximation to ϕ_{tilt}. These solutions have a small variance. The other solutions, however, have a relatively large variance. This suggests the following technique for determining an estimate ϕ_{etilt} of ϕ_{tilt}. Equation (4.62) is solved at, say, n far field sample points, yielding n pairs of solutions for ϕ_{tilt}. Out of the total of $2n$ solutions, the set of n solutions having the smallest variance is found. Remember that this set of n solutions must contain one solution from each pair of solutions. The value of ϕ_{etilt} is then set to the average value of this set of n solutions. The example presented below demonstrates a graphical procedure for determining the set of solutions which has the smallest variance.

The difference between $A_{\mathrm{m}}^{\times}(u, v)$ and $|F_{\mathrm{ent}}^{\times}|$ is dominantly due to either noise or the

$\phi_{\text{tilt}} F_{\text{ent}}$ term. The noise level is likely to be significantly less than $|\phi_{\text{tilt}} F_{\text{ent}}|$ at sample points for which

$$\frac{A_{\text{m}}^{\text{x}}(u,v) - |F_{\text{ent}}^{\text{x}}(u,v)|}{A_{\text{m}}(0,0)} > \gamma \Gamma_{\text{ran}} \tag{4.63}$$

where γ is a real number whose value is much greater than unity. Therefore, (4.62) should only be solved for sample points at which (4.63) holds. Note that this implies that $A_{\text{m}}^{\text{x}}(u,v)$ need only be sampled at such points.

An example is now presented which first illustrates the composite algorithm being applied to data in the presence of depolarization. The example then illustrates the above described technique for determining the feed tilt. The example concludes by simulating the correction of not only the geometrical defects, but also the feed tilt.

The particular model invoked in this section is an extension of the basic model and is defined by (cf. (4.47))

$$\begin{aligned}
&\text{design 2;} \qquad \Omega_{\text{pan}} = 0.019; \quad \psi_{\text{pan}} = 1.0 \text{ rad.;} \\
&\psi_{\text{quad}} = 1.0 \text{ rad.;} \quad \tau_{\text{ran}} = 0.01; \\
&\Gamma_{\text{ran}} = 0.001 = -60 \text{ dB;} \qquad \Gamma_{\text{off}} = 0.002 \\
&\tau_{\text{xpol}} = 0.03, \qquad \phi_{\text{tilt}} = 0.02
\end{aligned} \tag{4.64}$$

A cut through the design cross polar far field amplitude pattern $|F_{\text{d}}^{\text{x}}(u,v)|$ is shown in Figure 4.44(a). Because $|F_{\text{d}}^{\text{x}}(u,v)|$ vanishes along the u and v axes, but not elsewhere in the u,v plane, the cut is taken along a diagonal in the u,v plane. Position along this diagonal is identified by the parameter ν which is defined by (cf. (4.8))

$$\nu = \sqrt{2}\, u = -\sqrt{2}\, v \tag{4.65}$$

For comparison with $|F_{\text{d}}^{\text{x}}(u,v)|$, a cut through $|F_{\text{d}}(u,v)|$ along the same diagonal is also depicted in Figure 4.44(a). Note that $|F_{\text{d}}^{\text{x}}(0,0)| = 0$ thereby implying that there is no depolarization at the centre of the radiation pattern.

The value of ϕ_{tilt} in (4.64) is equivalent to a feed tilt angle of 1.15°. The actual cross polar and copolar far field amplitude patterns are depicted in Figure 4.44(b). The ratio $|F_{\text{a}}^{\text{x}}(0,0)|/|F_{\text{a}}(0,0)|$ is -34 dB.

To generate an estimate $f_{\text{e}}(x,y)$ of the copolar aperture field distribution, the composite algorithm is applied to $A_{\text{m}}(u,v)$ and $|f_{\text{d}}(x,y)|$. Note, from (4.19), that the non-zero value of ϕ_{tilt} implies that $|f_{\text{d}}(x,y)|$ is a less accurate estimate of $|f_{\text{a}}(x,y)|$ than it would have been had ϕ_{tilt} been zero. However, comparing the values of $\mathcal{E}^{\text{ap}}$, E_{m} and E_{c}, which are displayed in Figure 4.45, with those for the basic model (Fig. 4.25) shows that the convergence properties of the composite algorithm are only slightly affected by the non-zero value of ϕ_{tilt}.

Now that $f_{\text{e}}(x,y)$ is generated, an estimate for ϕ_{tilt} can be determined. Figure 4.46 depicts the pairs of approximate solutions for ϕ_{tilt}, computed by solving (4.62) for the centre nine far field sample points. It can be seen that the solutions, one from each pair, which are indicated by crosses, have a smaller range of values than the remaining solutions. Therefore, it is expected that the solutions indicated by the crosses are approximations to the correct value of ϕ_{tilt}. There are six far field sample points at which (4.63) is satisfied when $\gamma = 10$. The average value of the solutions indicated by the six crosses nearest the right hand edge of Figure 4.46, corresponding to these six sample points, is 0.0189. This average value is taken to be the estimate ϕ_{etilt} of ϕ_{tilt} and can be compared to the actual value 0.02 of ϕ_{tilt}.

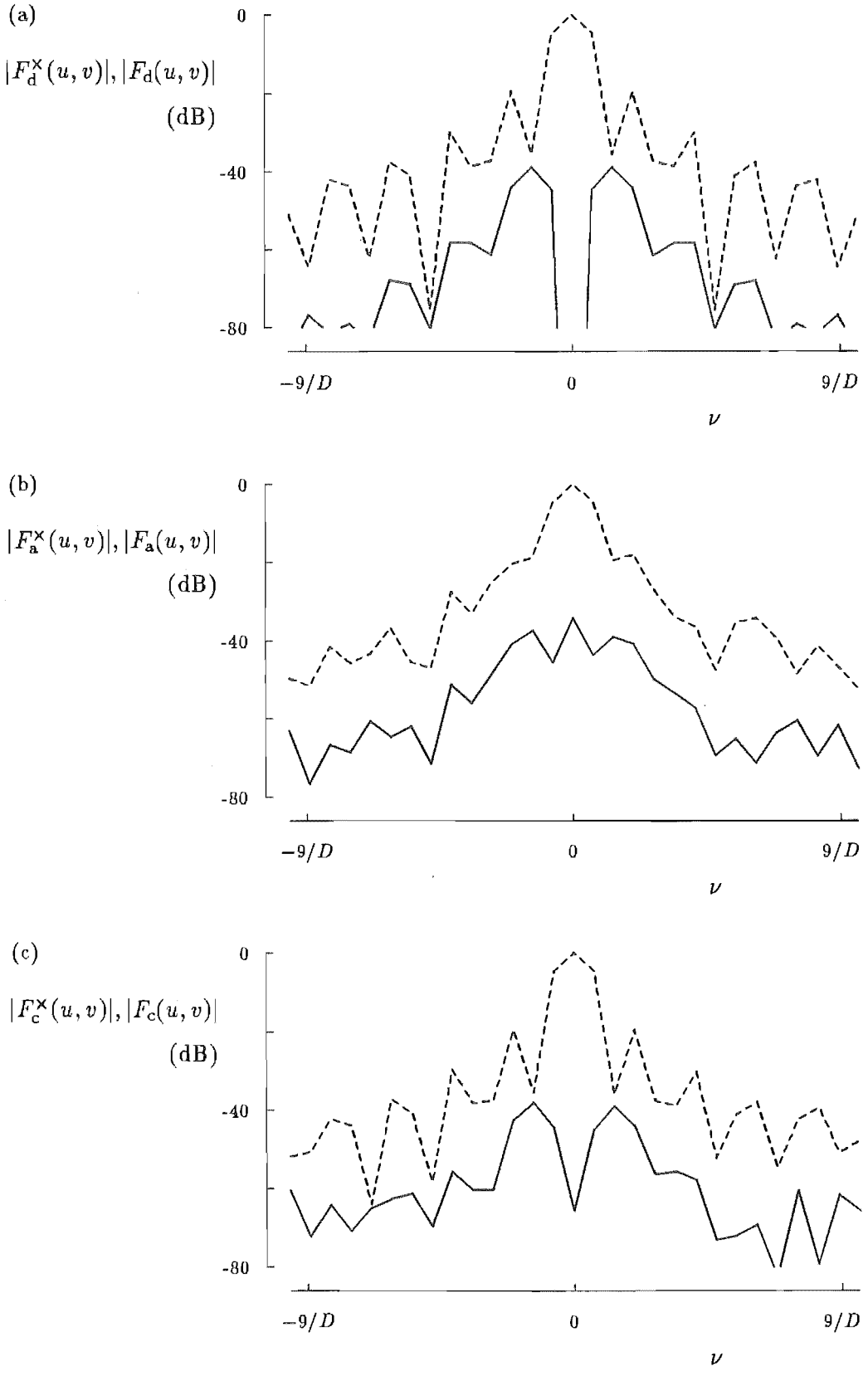


Figure 4.44 Cuts through the copolar (dashed curves) and cross polar (solid curves) far field amplitude patterns for the example, presented in Section 4.9.1, of estimating the feed tilt angle: (a) design amplitude patterns; (b) actual amplitude patterns; (c) corrected amplitude patterns. The variable ν is defined by (4.65).

Following the reasoning outlined in Section 4.3, but extending it to include the correction of feed tilt, the corrected aperture field distributions are here modelled by (cf. (4.25))

$$\begin{aligned} f_c(x,y) &= [f_a(x,y) - \phi_{etilt}\, f_a^{\times}(x,y)]e^{-j[\text{phase}\{f_e(x,y)\}-\psi_0]} \\ f_c^{\times}(x,y) &= [f_a^{\times}(x,y) + \phi_{etilt}\, f_a(x,y)]e^{-j[\text{phase}\{f_e(x,y)\}-\psi_0]} \end{aligned} \tag{4.66}$$

where it is assumed that $f_e(x,y)$, instead of $\tilde{f}_e(x,y)$, is the estimate of $f_a(x,y)$ and where, as explained in Section 4.3, the value of the real number ψ_0 is arbitrary.

Note that, in the example presented here, correcting the actual tilt ϕ_{tilt} of the feed on the basis of ϕ_{etilt} leaves a residual feed tilt, in the corrected antenna, of 0.0011 which is equivalent to a feed tilt angle of 0.063°. The corrected copolar and cross polar far field amplitude patterns are depicted in Figure 4.44(c). Note that the depolarization at the centre of this radiation pattern, which is −65.7 dB, is significantly less than for the actual radiation pattern depicted in Figure 4.44(b).

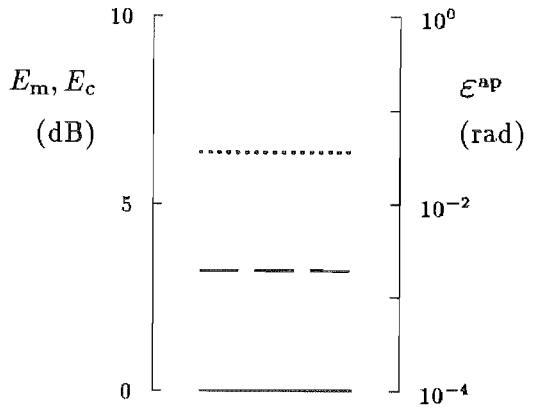


Figure 4.45 Errors for the depolarization example. The values of $\mathcal{E}^{ap}$ (dotted curve), E_m (dashed curve) and E_c (solid curve) pertain to the composite algorithm applied to the model defined by (4.64).

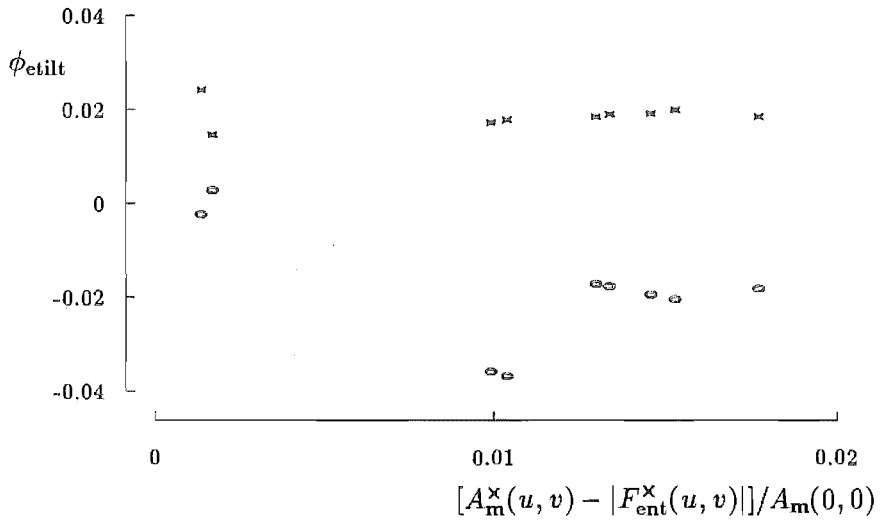


Figure 4.46 Solutions for ϕ_{tilt} obtained by solving (4.62). For each of the centre 9 far field sample points, the two solutions for ϕ_{tilt} are plotted against the value of $[A^{\times}_{\text{m}}(u,v) - |F^{\times}_{\text{ent}}(u,v)|]/A_{\text{m}}(0,0)$, which appears on the left side of (4.63). For each pair of solutions the larger one is is indicated by a cross, while the smaller one is indicated by an ellipse.

4.9.2 Aperture amplitude estimation

Most of the discussion in this chapter concentrates on the ability of the modified Gerchberg-Saxton algorithm, in the form of the composite algorithm, to accurately estimate the copolar aperture field phase distribution. This has been because, as intimated in Section 3.2.1, the most common geometrical defects cause deviations in the phase, but not the amplitude, of the copolar aperture field distribution. However, the composite algorithm generates an estimate of the amplitude, as well as the phase, of the copolar aperture field distribution. Knowing the copolar aperture field amplitude distribution $|f_{\text{a}}(x,y)|$ can aid with diagnosing reasons for an antenna not meeting its specifications. For example, if $|f_{\text{a}}(x,y)|$ is more tapered than $|f_{\text{d}}(x,y)|$ then the implication is that the feed has a narrower beamwidth than it was designed to have. The extent of scattering or blockage from struts can also be determined from $|f_{\text{a}}(x,y)|$. Furthermore, as pointed out by Bennett *et al.* [1976] in a complex holography context, knowing the full copolar aperture field distribution enables one to compute the radiation pattern at any distance from the antenna.

In the same way that $\mathcal{E}^{\text{ap}}$ is a measure of how accurately phase$\{f_{\text{e}}(x,y)\}$ approximates the phase of the image-form of $f_{\text{a}}(x,y)$, a measure of the accuracy to which $|f_{\text{e}}(x,y)|$ approximates the amplitude of the image-form of $f_{\text{a}}(x,y)$ is given by the *aperture amplitude error* $\mathcal{E}^{\text{aa}}$, which is defined to be either $\mathcal{E}^{\text{aa}}_a$ or $\mathcal{E}^{\text{aa}}_b$ where (cf. (4.21))

$$\mathcal{E}_a^{aa} = \left[\frac{\iint_{S^{f_d}} [|f_e(x,y)| - |f_a(x,y)|]^2 \, dx\, dy}{\iint_{S^{aper}} dx\, dy} \right]^{1/2}$$
$$\mathcal{E}_b^{aa} = \left[\frac{\iint_{S^{aper}} \left[|\tilde{f}_e(x,y)| - |f_a(x,y)|\right]^2 \, dx\, dy}{\iint_{S^{f_d}} dx\, dy} \right]^{1/2} \tag{4.67}$$

The reason for having a choice of values for $\mathcal{E}^{aa}$ is because, as discussed in Section 4.3, either $f_e(x,y)$ or $\tilde{f}_e(x,y)$ is the estimate of $f_a(x,y)$. The value of $\mathcal{E}^{aa}$ is taken to be $\mathcal{E}_a^{aa}$ if $\mathcal{E}^{ap}$ equals either $\mathcal{E}_a^{ap}$ or $\mathcal{E}_b^{ap}$. However, the value of $\mathcal{E}^{aa}$ is taken to be $\mathcal{E}_b^{aa}$ if $\mathcal{E}^{ap}$ equals either $\mathcal{E}_c^{ap}$ or $\mathcal{E}_d^{ap}$. As is true for $\mathcal{E}^{ap}$, $\mathcal{E}^{aa}$ cannot be computed in real-world situations. Note that, whereas $\mathcal{E}^{ap}$ is calculated over S^{f_d}, $\mathcal{E}^{aa}$ is calculated over S^{aper} so that it is affected when $f_e(x,y)$ does not properly predict the subreflector blockage.

Another estimate of $|f_a(x,y)|$ is $|f_d(x,y)|$. It is therefore of interest to know which of $|f_d(x,y)|$ and $|f_e(x,y)|$ is the better estimate of $|f_a(x,y)|$. To find this out, the value of $\mathcal{E}^{aa}$ can be compared with what is here called the *aperture data error* $\bar{\mathcal{E}}^{aa}$, which is the rms difference in amplitude between $f_a(x,y)$ and $f_d(x,y)$. $\bar{\mathcal{E}}^{aa}$ is defined by (cf. (4.67))

$$\bar{\mathcal{E}}^{aa} = \left[\frac{\iint_{S^{f_d}} [|f_d(x,y)| - |f_a(x,y)|]^2 \, dx\, dy}{\iint_{S^{aper}} dx\, dy} \right]^{1/2} \tag{4.68}$$

Unlike with the definition of $\mathcal{E}^{aa}$, the definition of $\bar{\mathcal{E}}^{aa}$ does not involve possible two choices for the aperture field amplitude because $|f_d(x,y)|$ for designs 1 and 2 are both circularly symmetric.

To illustrate the values of $\mathcal{E}^{aa}$ to which the composite algorithm converges when applied to a variety of particular models, $\mathcal{E}^{aa}$ and $\bar{\mathcal{E}}^{aa}$ are calculated for all the runs represented in subfigures (a) of Figures 4.31 to 4.36. Recall from the discussion in Section 4.8.1 that these runs involve models which can be defined by design 1, a quadratic aperture phase deviation and either an aperture amplitude deviation or far field measurement noise. The values of $\mathcal{E}^{aa}$ and $\bar{\mathcal{E}}^{aa}$ for these runs are depicted in Figure 4.47. Note that for the models invoked for Figure 4.47(c), $|f_a(x,y)| = |f_e(x,y)|$ so that $\bar{\mathcal{E}}^{aa} = 0$. It can be seen from Figure 4.47 that, in general, the composite algorithm converges to a higher value of $\mathcal{E}^{aa}$ the higher are the values of τ_{quad}, τ_{ran} or Γ_{ran}. The discussion in Section 4.8.4 indicates that there is a similar trend for $\mathcal{E}^{ap}$, implying that the composite algorithm performs worse the more inexact the data to which the algorithm is applied.

By comparing the values of $\mathcal{E}^{aa}$ with $\bar{\mathcal{E}}^{aa}$ represented in Figure 4.47(a) it can be observed that, for all of the models with a quadratic aperture amplitude deviation, $|f_e(x,y)|$ is a better estimate of $|f_a(x,y)|$ than is $|f_d(x,y)|$. However, Figure 4.47(b) shows that $|f_e(x,y)|$ is often no better an estimate of $|f_a(x,y)|$ than is $|f_d(x,y)|$ when the aperture amplitude deviation is noisy. These findings are more graphically illustrated in Figures 4.48 and 4.49, which depict the versions of $|f_a(x,y)|$, $|f_d(x,y)|$ and $|f_e(x,y)|$ corresponding to the large dots appearing in Figures 4.47(a) and (b) respectively.

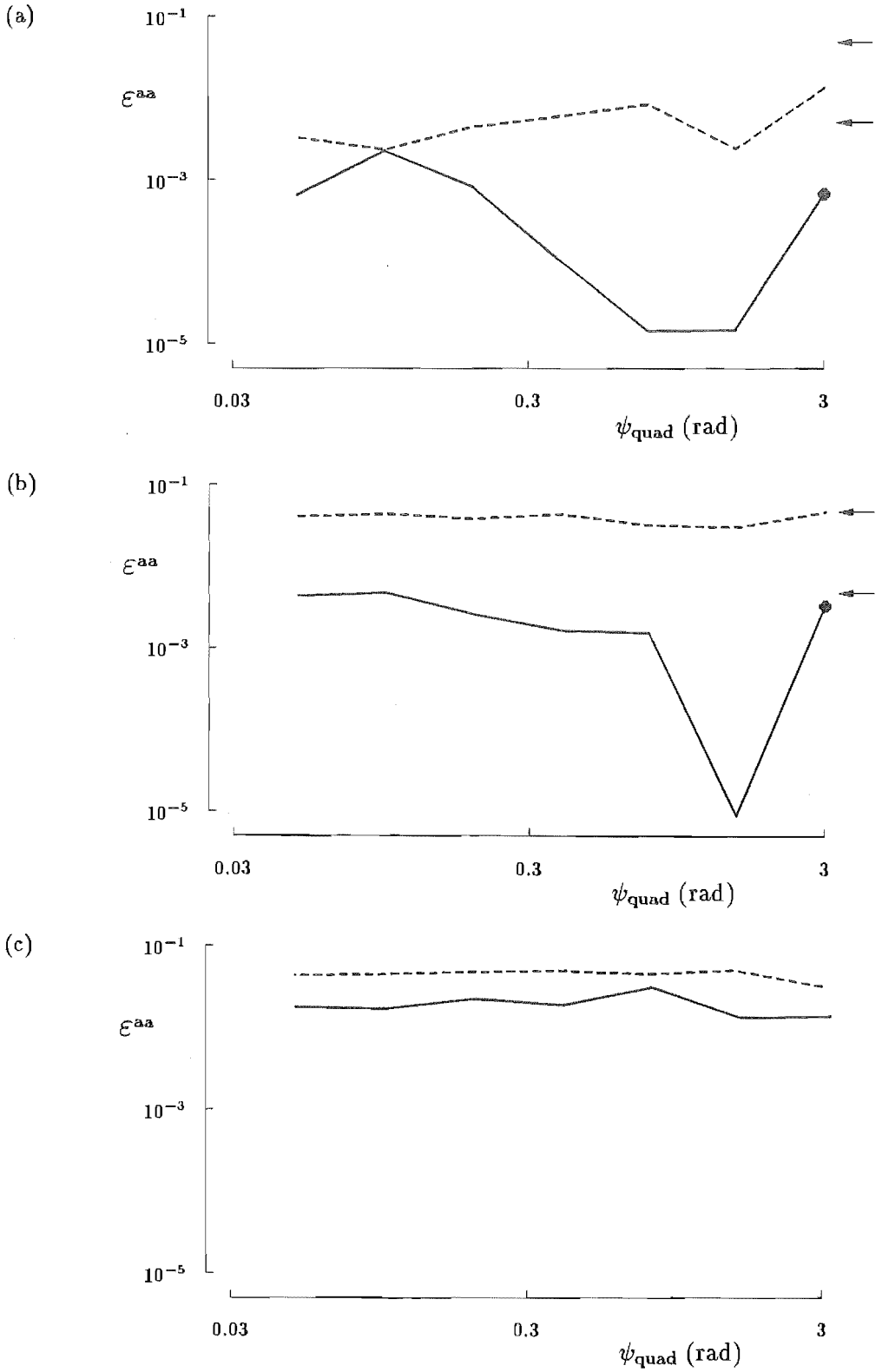


Figure 4.47 Accuracy to which the estimated copolar aperture field amplitude distribution approximates the amplitude distribution of the image-form of the actual field. The particular models invoked for the graphs are defined by design 1, $\psi_{quad} = \{0.05, 0.1, \ldots \text{or } 32\}$ rad. and either (a) $r_{quad} = 0.01$ (solid curve), $r_{quad} = 0.1$ (dashed curve), (b) $r_{ran} = 0.005$ (solid curve), $r_{ran} = 0.05$ (dashed curve), (c) $\Gamma_{ran} = -60$ dB (solid curve) or $\Gamma_{ran} = -50$ dB. In each of the top two graphs, the upper arrow indicates the value of the aperture data error $\bar{\mathcal{E}}^{aa}$ for the models invoked for the dashed curve, while the lower arrow indicates the value of $\bar{\mathcal{E}}^{aa}$ for the models invoked for the solid curve.

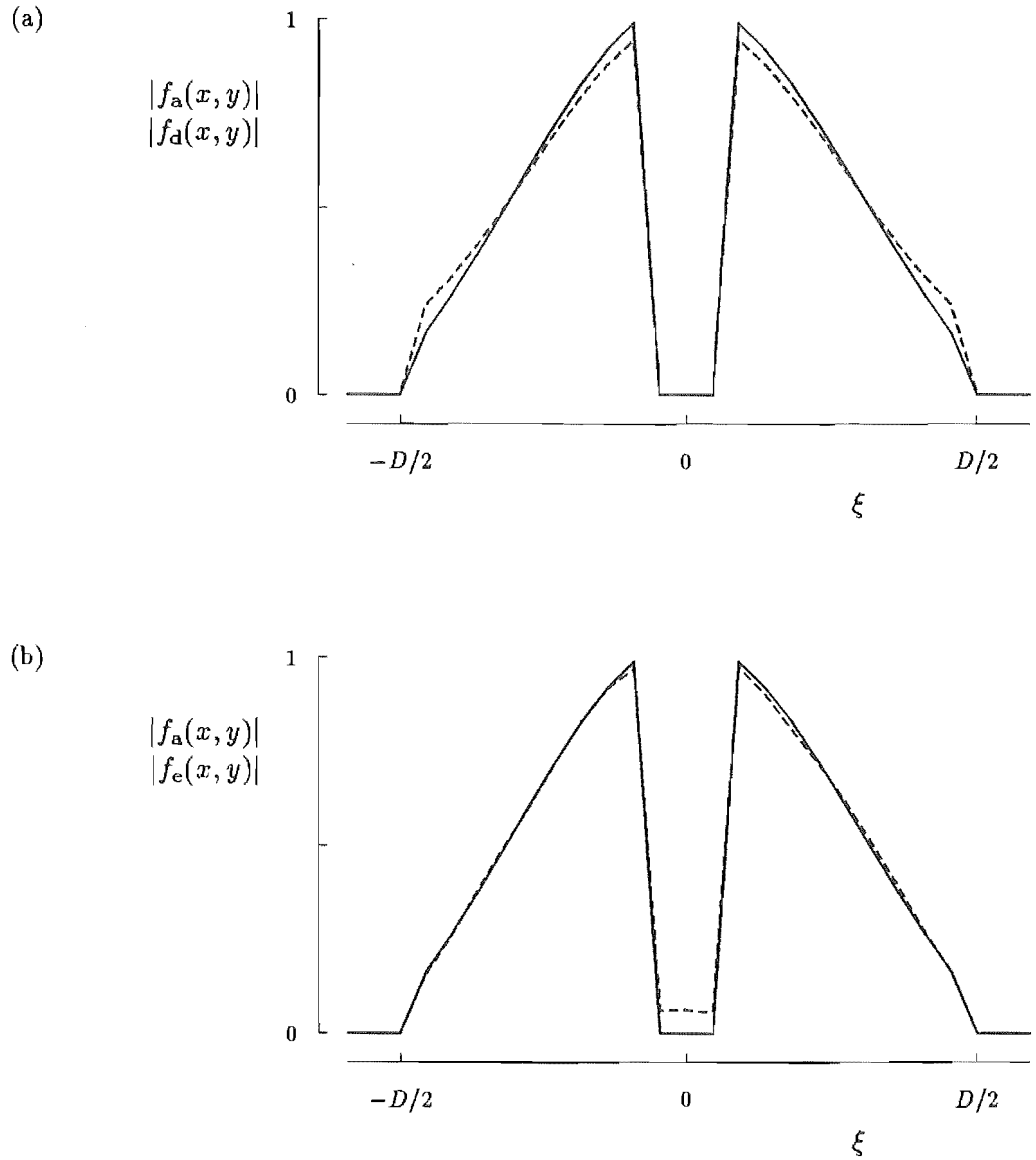


Figure 4.48 Comparison of the actual, design and estimate copolar aperture field amplitude distributions for the composite algorithm applied to data generated by the particular model defined by design 1, $\psi_{\text{quad}} = 3.2$ rad. and $r_{\text{quad}} = 0.1$: (a) $|f_a(x,y)|$ (solid curve) and $|f_d(x,y)|$ (dashed curve); (b) $|f_a(x,y)|$ (solid curve) and $|f_e(x,y)|$ (dashed curve). The variable ξ, which identifies position along a diagonal in the aperture plane, is defined in (4.8).

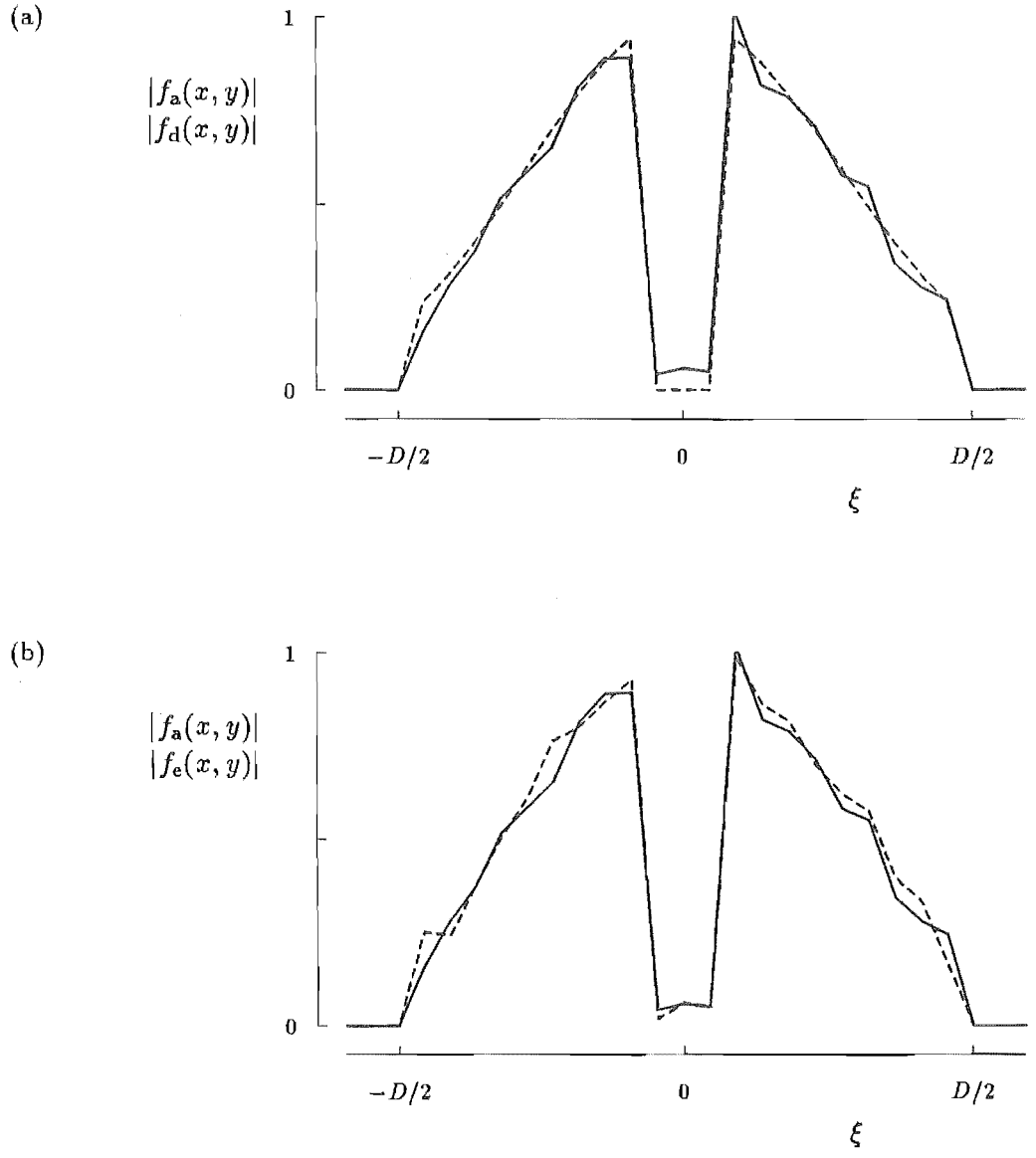


Figure 4.49 Comparison of the actual, design and estimate copolar aperture field amplitude distributions for the composite algorithm applied to data generated by the particular model defined by design 1, $\psi_{\text{quad}} = 3.2$ rad. and $\tau_{\text{ran}} = 0.05$: (a) $|f_a(x,y)|$ (solid curve) and $|f_d(x,y)|$ (dashed curve); (b) $|f_a(x,y)|$ (solid curve) and $|f_e(x,y)|$ (dashed curve). The variable ξ, which identifies position along a diagonal in the aperture plane, is defined in (4.8).

4.10 SUMMARY

This chapter describes the modified Gerchberg-Saxton algorithm, which is a generalized algorithm for generating an estimate of the copolar aperture field distribution and requires only a single measurement of the copolar far field amplitude pattern to be made. One form of the modified Gerchberg-Saxton algorithm, called the composite algorithm, is examined in considerable detail and is evaluated by applying it to data generated from computer models.

The purpose of the modified Gerchberg-Saxton algorithm is to generate an estimate $f_e(x,y)$ of the actual copolar aperture field distribution $f_a(x,y)$. The inputs to the algorithm are the design copolar aperture field amplitude distribution $|f_d(x,y)|$ and the measured copolar aperture amplitude pattern $A_m(u,v)$. $|f_d(x,y)|$ is either obtained from design data or is taken to be a guess of the likely form of $|f_a(x,y)|$.

The modified Gerchberg-Saxton algorithm attempts to find a copolar aperture field distribution $f_e(x,y)$ which is constrained to be zero where $|f_d(x,y)|$ is zero and whose corresponding copolar far field amplitude pattern $|F_e(u,v)|$ is constrained to approximate $A_m(u,v)$. The algorithm starts with a copolar aperture field distribution whose amplitude equals $|f_d(x,y)|$ and whose phase is random. It then iterates between the aperture plane and the far field plane, applying the constraints in each plane. It is shown in Section 3.4.2 that, provided the algorithm has converged sufficiently, in the sense that both constraints are approximately met, $f_e(x,y)$ is an estimate of the image-form of $f_a(x,y)$. This means that either $f_e(x,y)$ or $\tilde{f}_e(x,y)$, which is the conjugate reflection of $f_e(x,y)$, is the estimate of $f_a(x,y)$. This ambiguity can be resolved in a number of ways which are discussed in Section 4.1.2.

There are a number of different ways in which the above-mentioned constraints can be applied. I have found that constraints requiring the copolar aperture field amplitude distribution to equal $|f_d(x,y)|$ can often enable the algorithm to converge where it might otherwise have stagnated. Constraints involving $|f_d(x,y)|$ are utilized by the CC and HIO algorithms, which are forms of the modified Gerchberg-Saxton algorithm based on, respectively, a variant of the Gerchberg-Saxton algorithm (Sec. 3.4.3.2) and Fienup's hybrid input-output algorithm (Sec. 3.4.3.3). Of the several forms of the modified Gerchberg-Saxton algorithm with which I have worked, the one whose convergence characteristics I have found to be best overall is the composite algorithm. This algorithm consists of running the CC and HIO algorithms each three times and choosing the best of the 6 generated distributions $f_e(x,y)$. The best $f_e(x,y)$ is chosen to be the one for which $|F_e(u,v)|$ is closest to $A_m(u,v)$. It is therefore the composite algorithm which is evaluated in this chapter.

Recall that the reason for wanting to generate $f_e(x,y)$ is to be able to identify and then correct geometrical defects of a high gain reflector antenna. Section 3.2 outlines how these defects can be located from the information contained in phase$\{f_e(x,y)\}$. After it has been corrected, the aperture field of the antenna is here called the corrected copolar aperture field distribution $f_c(x,y)$.

In order to apply the constant correction algorithm to a wide variety of data, a generalized computer model has been developed to generate such data. The model simulates the copolar aperture field distributions and copolar far field radiation patterns of an antenna and the process of measuring the amplitude pattern. A copolar aperture field distribution and its corresponding copolar far field pattern are related by Fourier transformation. Two different distributions for $f_d(x,y)$ are modelled: one is typical of a prototype Cassegrain antenna while the other is associated with a shaped Cassegrain

antenna. The actual copolar aperture field $f_a(x,y)$ can differ from $f_d(x,y)$ for many reasons including: aperture phase deviations caused by a combination of displaced panels and defocused feed or subreflector; aperture amplitude deviations due to a feed pattern with a different beamwidth than specified by the design; noise-like interference in the copolar aperture field due to scattering from struts. All of these effects are simulated in the generalized model. The model also simulates measurement inaccuracies such as those caused by measurement noise, calibration inaccuracy and truncation caused by making measurements over too small a region of the u,v plane. The effect of all of these measurement inaccuracies is to cause $A_m(u,v)$ to differ from $|F_a(u,v)|$.

One of the advantages of applying the modified Gerchberg-Saxton algorithm to computer generated data is that the performance of the algorithm can be assessed by comparing $f_e(x,y)$ with $f_a(x,y)$, which cannot of course be done when the algorithm is applied to real-world data (because there is no way of deducing with arbitrary accuracy the copolar aperture field distribution corresponding to a measured copolar far field amplitude pattern). Therefore, in this chapter, various measures for determining the accuracy and usefulness of $f_e(x,y)$, generated by the modified Gerchberg-Saxton algorithm, have been developed. The aperture phase error $\mathcal{E}^{ap}$ is a direct measure of the accuracy to which phase$\{f_e(x,y)\}$ approximates the phase of the the image-form of $f_a(x,y)$. The corrected envelope error E_c provides an indication of whether or not the corrected copolar far field amplitude pattern $|F_c(u,v)|$ meets its specifications. The calculation of E_c involves modelling $f_c(x,y)$ by simulating the process of correcting, according to the information within phase$\{f_e(x,y)\}$, the geometrical defects.

In Section 4.8 the composite algorithm is applied to data obtained from a wide variety of particular models. It is encouraging that the convergence of the composite algorithm, with respect of E_c, does not, in general, appear to depend on the level of aperture phase deviation. However, the convergence of the composite algorithm does depend upon the accuracy of the data to which it is applied. The results of the computer simulations are summarized in Section 4.8.4.

Provided $A_m(u,v)$ is oversampled by a factor of at least two, and if the modified Gerchberg-Saxton algorithm converges well enough, it is guaranteed to converge towards the image-form of $f_a(x,y)$. However, it appears from the results presented in Section 4.1.1 that, even when $A_m(u,v)$ is oversampled by a factor of only 1.7, the composite algorithm usually generates a useful estimate of the image-form of $f_a(x,y)$. Smoothing $A_m(u,v)$ in an effort to remove the high spatial frequency component of the measurement noise does not appear to help the convergence of the composite algorithm. Three different methods have been developed for dealing with the truncation of $A_m(u,v)$ due to the range of angles over which the far field pattern is measured being too small. Of these methods, the most successful, on the basis of the results presented in Section 4.7.3, is an adaptation of the composite algorithm which enables it to extrapolate the far field data while performing phase retrieval.

Although the purpose for which the composite algorithm was developed is to generate an accurate estimate of the phase of $f_a(x,y)$, the algorithm can be utilized for more than just that. For example, the amplitude of $f_e(x,y)$, generated by the composite algorithm, can often be a more accurate estimate of $|f_a(x,y)|$ than $|f_d(x,y)|$ is. It is also possible to utilize phase$\{f_e(x,y)\}$ to estimate various causes of depolarization.

CHAPTER 5

EXPERIMENTAL VERIFICATION OF MODIFIED GERCHBERG-SAXTON ALGORITHM USING AN ACOUSTIC ANTENNA

In the previous chapter, the modified Gerchberg-Saxton algorithm is applied to data generated from a computer model. Any such computer model can never simulate exactly the data that would be obtained by measuring a real-world radiation pattern. Therefore, to provide more realistic data for the algorithm to operate upon, the radiation pattern of an acoustic antenna has been measured. This acoustic experiment is described in this chapter.

Ideally, the experiment would have involved a high gain radio antenna, since this is the type of antenna that is studied in this thesis. However, neither such an antenna, nor a suitably measured radiation pattern of such an antenna, was available to me. The decision to instead utilize an acoustic antenna was made for the following reasons. In order to test the modified Gerchberg-Saxton algorithm it is necessary to use sources which are related to their radiation patterns by Fourier transformation. This relationship occurs for acoustic antennas, as is discussed in Section 5.1. The speed of sound in air is many orders of magnitude slower than the speed of radio waves in air. This means that radio waves of gigahertz frequencies have the same wavelength as sound waves of audio frequencies. It is much easier and cheaper to design and construct electronic hardware to operate at these lower frequencies than at the higher frequencies. Furthermore, use can be made of the expertise on acoustic systems which is available in the Electrical and Electronic Engineering Department at the University of Canterbury.

Ott and Rice [1986] describe an acoustic simulator of high gain radio antennas comprising an array of 3348 speakers, each feed by a signal whose phase and amplitude is under computer control. The antenna involved in the experiments described in this chapter is similar, but is not nearly as ambitious: it comprises only nine speakers. It does meet its purpose, however, which is to produce a variety of amplitude patterns each corresponding to a different aperture field distribution.

Section 5.1 introduces a mathematical model of the propagation of acoustic waves and compares acoustic wave theory with electromagnetic wave theory. The acoustic antenna and the apparatus utilized to measure the antenna's radiation pattern are described in Section 5.2. Far field amplitude data from two different aperture field distributions have been recorded. Section 5.3 presents the results of applying the modified Gerchberg-Saxton algorithm to these data.

5.1 ACOUSTIC WAVES

The theory of acoustic waves is briefly outlined in this section. Starting with the basic equations which describe acoustic wave propagation, the Fourier transform relationship

between the aperture field distribution and the far field pattern of an acoustic antenna (or transducer) is established. Because it roughly parallels the development of electromagnetic wave theory, there are many references in this section to the theory presented in Sections 1.1, 2.1.2 and 2.1.3.

Consider a compressible fluid. A *particle* of the fluid is a volume of fluid whose dimensions are large compared to those of the molecules, so that the fluid can be thought of as a continuous medium, yet small enough for the variations of the acoustic quantities to be infinitesimal throughout the particle [Kinsler *et al.*, 1982, p. 99]. An acoustic wave is initiated by disturbing the position of one or more such particles. This disturbance is transmitted from one particle to the next and so on throughout the entire medium. During its motion, each particle is subjected to successive compressions and expansions, with accompanying variations in pressure and volume. This process is characteristic of the propagation of an acoustic wave [Rossi, 1988, p. 6].

An acoustic wave can be described by the way the acoustic quantities, such as pressure or density, vary with spatial position $\mathbf{r}$ and time t. Arbitrary time dependence is assumed for the three acoustic quantities listed below. The following definitions are taken from Rossi [1988, Sec. 1.2.3]:

1. The *particle velocity* $\mathbf{v}(\mathbf{r}, t)$:

$$\mathbf{v}(\mathbf{r}, t) = \frac{\partial \xi(\mathbf{r}, t)}{\partial t} \tag{5.1}$$

   where $\xi(\mathbf{r}, t)$ is the particle displacement from its equilibrium position.

2. The *acoustic pressure* $p(\mathbf{r}, t)$, which indicates the local variation in pressure:

$$p(\mathbf{r}, t) = \mathcal{P}(\mathbf{r}, t) - \mathcal{P}_\mathrm{a} \tag{5.2}$$

   where $\mathcal{P}(\mathbf{r}, t)$ is the instantaneous pressure and $\mathcal{P}_\mathrm{a}$ is the ambient pressure, which is what exists in the absence of any acoustic disturbance.

3. The *condensation* $s(\mathbf{r}, t)$, which is the relative variation of mass density:

$$s(\mathbf{r}, t) = \rho(\mathbf{r}, t) - \rho_\mathrm{a} \tag{5.3}$$

   where $\rho(\mathbf{r}, t)$ is the instantaneous mass density and ρ_a is the ambient density, which is what the mass density would be in the absence of any acoustic disturbance.

Throughout the remainder of this chapter, harmonic time dependence is assumed. Recall from Section 1.1.1 that a quantity, for example acoustic pressure, which varies harmonically with time is denoted by $p(\mathbf{r})$, because the exponential factor $e^{j\omega t}$ is understood. By analogy with (1.1), $p(\mathbf{r})$ is defined in terms of a time harmonic $p(\mathbf{r}, t)$ through

$$p(\mathbf{r}, t) = \mathrm{real}\left\{p(\mathbf{r})e^{j\omega t}\right\} \tag{5.4}$$

where ω is angular frequency. Similarly, harmonically time varying particle velocity and condensation are denoted $\mathbf{v}(\mathbf{r})$ and $s(\mathbf{r})$ respectively.

In the following analysis it is assumed that the fluid, through which the acoustic wave propagates, is not subject to external forces, such as gravity. Thus, $\mathcal{P}_\mathrm{a}$ and ρ_a are constant throughout the medium. The fluid is also assumed to be homogeneous, isotropic and perfectly elastic, meaning that there is no loss of energy in the form

of dissipation of thermal energy. The analysis is limited to waves of small amplitude (i.e. $s \ll 1$) thereby allowing the linearizing of equations. These assumptions are required to develop the simplest theory for acoustic waves in fluids. Kinsler *et al.* [1982, Sec. 5.1] note that this theory adequately describes most common acoustical phenomena. Sources of acoustic wave motion are disregarded here: only the propagation of the waves is analysed.

The equations describing acoustic waves are derived from basic physical laws which, for time harmonic wave motion, are [cf. Kinsler *et al.*, 1982, Secs. 5.3–5.5; Rossi, 1988, Secs. 1.2.5–1.2.7]

$$\begin{aligned} \nabla p &= -j\omega\rho_a \mathbf{v} \\ \nabla \cdot \mathbf{v} &= -j\omega s \\ s &= B^{-1} p \end{aligned} \tag{5.5}$$

where B is the bulk modulus of the fluid. The first equation of (5.5) is a local form of Newton's second law of motion: the force necessary to move a particle must be equal to the rate of change of its momentum. The second equation of (5.5) is called the continuity equation and is a local form of the principle of conservation of mass. The final equation of (5.5) is a local form of the compressibility law relating the pressure experienced by a particle to its mass density. Although B is assumed to be independent of p and s, it is dependent upon the ambient state of the fluid [Rossi, 1988, p. 12].

An acoustic wave is fully characterized by any one of the acoustic quantities p, $\mathbf{v}$ or s. Throughout this chapter, the acoustic pressure p is invoked. From it, $\mathbf{v}$ and s can be obtained via (5.5). The acoustic equations (5.5) can be transformed into a wave equation for acoustic pressure [cf. Rossi, 1988, Sec. 1.2.8; Kinsler *et al.*, 1982, Sec. 5.5]:

$$\nabla^2 p + k^2 p = 0 \tag{5.6}$$

where $k = \omega[\rho_a/B]^{1/2} = 2\pi/\lambda$ is the wave number. The speed of propagation of an acoustic wave is $c = (B/\rho_a)^{1/2}$. A typical value for c in air is 340 metres/second [Rossi, 1988, Sec. 1.2.17].

Figure 5.1 depicts all of space divided into two half-spaces V_1 and V_2, separated by an infinite plane surface S. The sources of an acoustic wave are located within V_1. The location of an arbitrary point on S is identified by the position vector $\mathbf{r}'$, at which the normal to S is indicated by $\hat{\mathbf{n}}$. It follows from the wave equation (5.6), that the acoustic pressure at any point $\mathbf{r}$ in V_2 is [Goodman, 1968, Sec. 3.4]

$$p(\mathbf{r}) = \frac{1}{4\pi} \int_S \left\{ \frac{\partial p}{\partial n} \psi - p \frac{\partial \psi}{\partial n} \right\} dS \tag{5.7}$$

where dS is an elemental area of S, $\partial \cdot / \partial n$ denotes the normal derivative and $\psi(\mathbf{r}')$ is a Green's function. The particular form of $\psi(\mathbf{r}')$ adopted here is [Goodman, 1968, p. 43]

$$\psi(\mathbf{r}') = \frac{e^{-jkr_d}}{r_d} - \frac{e^{-jkr_{dx}}}{r_{dx}} \tag{5.8}$$

where $r_d = |\mathbf{r} - \mathbf{r}'|$ and $r_{dx} = |\mathbf{r} - 2\mathbf{r} \cdot \hat{\mathbf{n}} - \mathbf{r}'|$. When (5.8) is substituted into (5.7), the latter equation simplifies to

$$p(\mathbf{r}) = \frac{j}{\lambda} \int_S p(\mathbf{r}') \frac{e^{jkr_d}}{r_d} \hat{\mathbf{n}} \cdot \hat{\mathbf{r}}_d \, dS \tag{5.9}$$

where $\hat{\mathbf{r}}_d$ is the unit vector in the direction of the vector $(\mathbf{r} - \mathbf{r}')$. Equation (5.9)

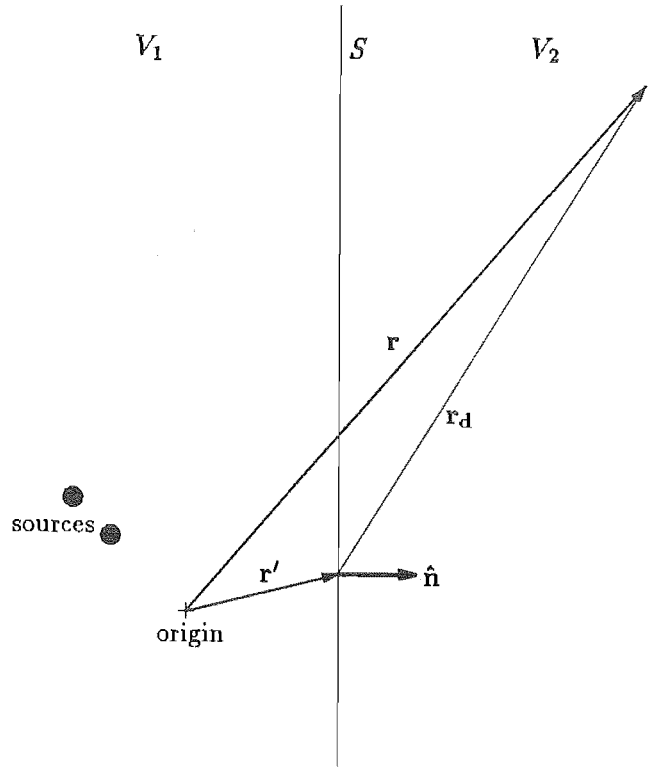


Figure 5.1 Geometry for the theory developed in Section 5.1.

therefore expresses the acoustic pressure, at any point in V_2, in terms of the acoustic pressure distribution over the plane S.

Following the reasoning presented in (2.1.2.2), the approximations to r_d and $\hat{\mathbf{r}}_d$ which are listed in (2.22) are valid when $\mathbf{r}$ is in the far field of the acoustic sources. Substituting (2.22) into (5.9) gives

$$p(\mathbf{r}) = \frac{je^{-jkr}}{\lambda r}\,\hat{\mathbf{n}}\cdot\hat{\mathbf{r}}\int_S p(\mathbf{r}')e^{jk\mathbf{r}'\cdot\hat{\mathbf{r}}}\,dS \tag{5.10}$$

Let the planar surface S coincide with the x, y plane of a Cartesian coordinate system x, y, z, so that $\hat{\mathbf{n}} = \hat{\mathbf{z}}$. The acoustic pressure distribution over this plane is the aperture field distribution, while the angular distribution of the acoustic pressure in the far field region is the far field pattern. Utilizing the definitions (2.27) and (2.28), the aperture field distribution and the far field pattern are respectively denoted by

$$\begin{aligned} p(x,y) &= p(x,y,0) = p(\mathbf{r}') \\ \acute{p}(u,v) &= p(R\lambda u, R\lambda v, R\lambda w) = p(\mathbf{r}) \end{aligned} \tag{5.11}$$

With the aid of (2.27) and (2.28), individual terms appearing in (5.10) can be simplified:

$$\begin{aligned} \hat{\mathbf{n}}\cdot\hat{\mathbf{r}} &= \lambda w(u,v) \\ jk\mathbf{r}'\cdot\hat{\mathbf{r}} &= j2\pi(xu+yv) \\ dS &= dx\,dy \end{aligned} \tag{5.12}$$

Substituting (5.11), (5.12) and (2.28) into (5.10) yields

$$\begin{aligned} \acute{p}(u,v) &= jw(u,v)\frac{e^{-jkR}}{R}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)e^{j2\pi(xu+yv)}\,dx\,dy \\ &= jw(u,v)\frac{e^{-jkR}}{R}\mathrm{FT}\{p(x,y)\} \end{aligned} \tag{5.13}$$

In order to maintain the notation employed in Chapter 4, the following definitions are introduced:

$$f(x,y) = p(x,y) \quad \text{and} \quad F(u,v) = \frac{1}{jw(u,v)}\frac{R}{e^{-jkR}}\acute{p}(u,v) \tag{5.14}$$

where $f(x,y)$ is the aperture field distribution and $F(u,v)$ is here called the *weighted far field pattern*. It follows from (5.13) and (5.14) that $F(u,v)$ is the Fourier transform of $f(x,y)$. Note that, although the notation invoked in Chapter 4 is also invoked in this chapter, the terminology is slightly different. For example, $f_{\mathrm{a}}(x,y)$ and $F_{\mathrm{a}}(u,v)$ are respectively referred to in this chapter as the actual aperture field distribution and the actual weighted far field pattern.

Acoustic waves are now compared with electromagnetic waves. Rossi [1988] notes that the basic time harmonic acoustic equations (5.5) and Maxwell's equations for time harmonic electromagnetic waves (1.9) are of similar type. They are both sets of algebraic relations linking quantities to themselves and to their derivatives with respect to space. Comparing (5.6) with the left side of the first equation of (1.14) reveals the similarity between the mathematical models of acoustic and electromagnetic wave propagation. The main difference between these two models is that acoustic waves are characterized by a scalar quantity (e.g. p), whereas electromagnetic waves are characterized by a vector quantity (e.g. $\mathbf{E}$). Note that the equation (5.13), relating $p(x,y)$ to $\acute{p}(u,v)$, has exactly the same form as, say, the first equation of (2.31) which relates $E_x(x,y)$ to $\acute{E}_x(u,v)$. This suggests that, provided the conditions required to develop (2.31) and (5.13) are met, components of $\mathbf{E}(\mathbf{r})$, such as $E_x(\mathbf{r})$, can be treated as scalar quantities which are independent of other components of $\mathbf{E}(\mathbf{r})$ [Silver, 1949, Sec. 4.1; Goodman, 1968, Sec. 3.1]. Note that the only formal difference between the definitions of $F(u,v)$ in (3.74) and (5.14) is the inclusion of the weighting term $1/w(u,v)$ in the latter. The reason for this difference is that the far field pattern for high gain radio antennas is negligible outside the small angle region (2.32). Within this region, $1/w(u,v)$ reduces effectively to λ. By contrast, the far field pattern, for the particular acoustic antenna described in this chapter, is significant outside the small angle region, thereby necessitating the inclusion of the $1/w(u,v)$ term in (5.14).

5.2 EXPERIMENTAL APPARATUS

The purpose of the experimental apparatus is to provide a measured weighted far field amplitude pattern to which the modified Gerchberg-Saxton algorithm can be applied. The measuring arrangement is such that the antenna, whose amplitude pattern is measured, transmits an acoustic wave. This conveniently allows the antenna to be constructed from speakers and its far field pattern to be measured with a microphone. In order to be able to apply the algorithm to a variety of data, the antenna can be configured to generate a variety of aperture field distributions. The antenna and the hardware used to drive it are described in Section 5.2.1. Cuts through the amplitude

pattern of the antenna are measured and plotted on a pen plotter. Section 5.2.2 describes the hardware employed to make these measurements. The data in the plots are mathematically transformed by a computer into a form suitable to be operated upon by the modified Gerchberg-Saxton algorithm. The computer software which permits this is outlined in Section 5.2.3.

5.2.1 The antenna

The antenna consists of a 3 by 3 array of speakers (one inch dome tweeter loudspeakers). The speakers are mounted so that they face the same direction with all their physical apertures coinciding with a common plane: the aperture plane of the antenna. The speakers are mounted as close as possible to each other. Their physical arrangement is indicated in Figure 5.2. The speakers cover a 0.225 m × 0.225 m area of the aperture plane. Each speaker is uniquely identified by a number from 1 to 9, as indicated in Figure 5.2.

To reduce the effects of interference from external sources of acoustic noise, the measurements are made in an anechoic chamber. Since the ambient temperature in this

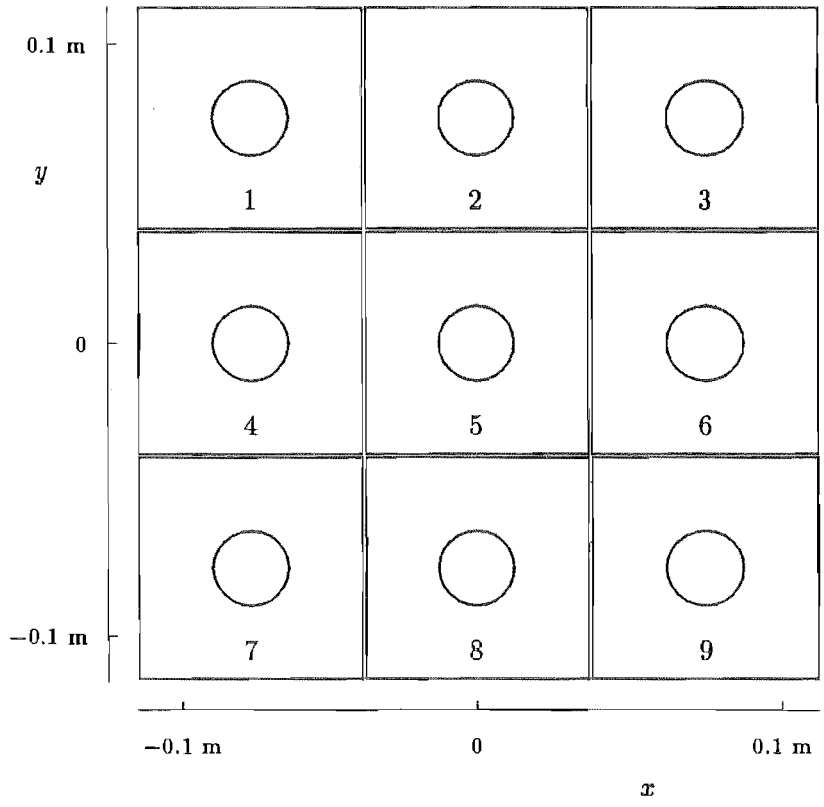


Figure 5.2 The acoustic antenna. The positions in the aperture plane of the nine speakers comprising the antenna are indicated. The circles indicate the perimeters of the physical apertures of the speakers. Each square indicates the perimeter of the flange which surrounds each speaker. Each speaker is identified by a number.

chamber is always close to $20°C$, the speed of acoustic wave propagation is assumed to be 340 m/s [Rossi, 1988, Sec. 1.2.17]. The antenna is operated at a frequency of 12.5 kHz which corresponds to a wavelength of $\lambda = 0.0272$ m. At this frequency, assuming aperture field to extend over a square area of 0.195 m × 0.195 m (cf. Sec. 5.3.4), the far field distance calculated from (1.20) is 2.88 m. The size of the anechoic chamber is such that the the furthest distance the measuring microphone can be from the antenna is 3.39 m. At this distance, the measuring microphone is therefore in the far field region of the wave radiated by the antenna.

The reason for choosing to operate at a frequency of 12.5 kHz is that it corresponds to a peak in the frequency response of the speakers and measurement equipment which were available to me. At such a frequency the signal to noise ratio of the measured signal is expected to be greater than at other frequencies. The peak with the next highest frequency is at 16.5 kHz, which would place the microphone closer than the far field distance. Any frequency lower than 12.5 kHz would have the negative effect of widening the radiation pattern.

The lining of the particular anechoic chamber available to me is such that, at 12.5 kHz, it reflects acoustic waves to an appreciable degree. Therefore, especially when the main beam of the antenna is pointed towards the floor, a reflected wave of significant amplitude interferes with the direct wave between the antenna and the microphone. To avoid this unwanted interference, the signal fed to the antenna is not a continuous, time harmonic signal, but is instead a series of tone bursts. Consider a single tone burst emitted from the antenna. Because the direct path between the antenna and the microphone is shorter than any indirect path (e.g. via a wall of the acoustic chamber) between the antenna and the microphone, the directly propagated tone burst arrives at the microphone before any reflected tone burst. The duration of each tone burst must be short enough for the microphone to receive the whole burst directly propagated from the speakers before it starts to receive the first of the reflected bursts. The transmitted bursts must be separated by a time interval sufficiently long to allow the acoustic energy associated with one burst to be absorbed by the walls of the chamber before the next burst is transmitted. The signal received by the microphone is only measured while the directly propagated tone burst is received. Consequently each tone burst can be treated as if it were part of a continuous time harmonic signal propagating without distortion or disturbance from the antenna to the measuring microphone.

A schematic of the hardware utilized for the experiment is shown in Figure 5.3. The hardware which drives the antenna consists of a waveform generator, a phase delay unit, a switch board and a set of power amplifiers. These components feed each speaker with its own a signal. The amplitude and phase of the tone within the tone burst can be adjusted for each individual speaker in the antenna. The hardware which drives the antenna is described in the following paragraphs.

The waveform generator generates two signals which are here called the reference signal and the tone burst signal. The reference signal is a continuous sinusoid with a frequency of 12.5 kHz. It is fed to the phase delay unit and the phase meter (Sec. 5.2.2). The reference signal is generated by an 8038 precision waveform generator configured as an audio oscillator [Intersil, 1981, p. 5-196]. The tone burst signal is derived from the reference signal by passing the latter through a voltage controlled switch. The control voltage for this switch is synchronized with the reference signal so that the switch is turned on when the reference signal passes through zero volts and is turned off after

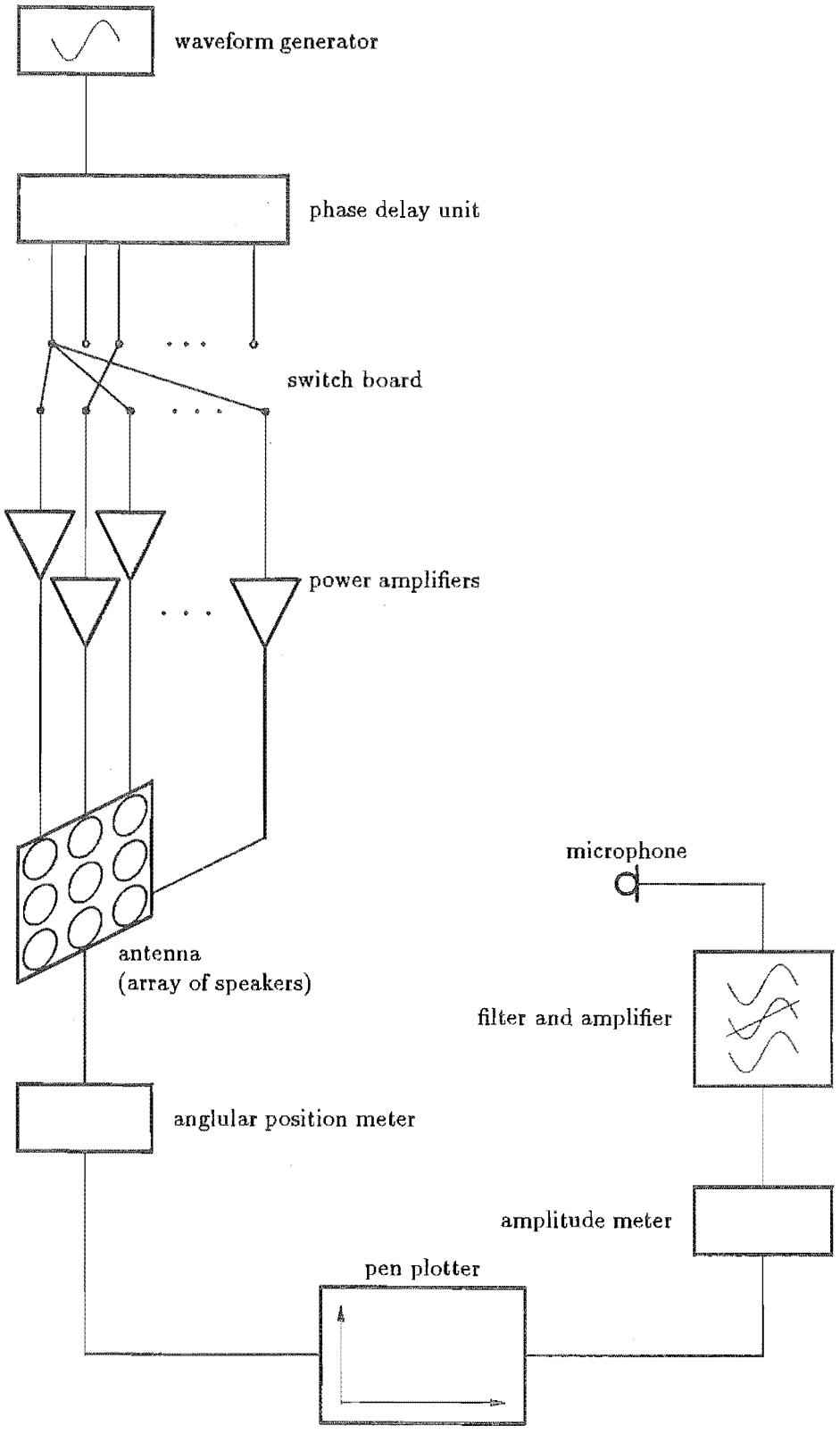


Figure 5.3 The experimental apparatus. The function of each of each piece of equipment is outlined in Sections 5.2.1 and 5.2.2.

a further 16 cycles of the reference signal. This turning on and off of the switch is repeated at 315 ms intervals.

The phase delay unit, to which the tone burst signal is fed, has 16 outputs, each of which is a time delayed version of the tone burst signal. The design of the delay unit is based around a TAD32 tapped analogue delay chip [Reticon, 1983, Chap. 5]. This chip is controlled by a signal whose frequency is 16 times that of the reference signal [National, 1981, p. 5-129], thereby ensuring that the phase delay between successive outputs of the phase delay unit is $\pi/8$ rad.

Each speaker is driven by a dedicated power amplifier. The amplitude of the signal fed to each speaker is determined by the gain of the corresponding power amplifier. The gain of each amplifier can be adjusted by a potentiometer. The switch board can be variably configured, with a system of plugs and sockets, to connect any of the 16 phase delay unit outputs to any number of the power amplifier inputs. In this way each power amplifier can be independently fed by any of the 16 phase delayed tone burst signals. Through the combination of the phase delay unit, the switch board, the power amplifiers and their gain settings, the antenna can be configured to produce a variety of aperture distributions, each one obtained by feeding the individual speakers with signals of different phase and amplitude.

5.2.2 Measurement hardware

The purpose of the measurement hardware is to measure and record the far field amplitude pattern of the antenna described in the previous section. The hardware for accomplishing this aim includes a mount which enables the antenna to rotate, a measurement microphone with its associated circuitry, an amplitude meter and a pen plotter. These components are indicated schematically in Figure 5.3 and are described in this section.

As intimated in Section 1.2.2, the far field pattern of an antenna is the angular distribution of a quantity, in this case acoustic pressure, which characterizes the far field. In order to measure the angular variation of acoustic pressure, the antenna is mounted so that it can rotate about two axes, while the position of the measurement microphone is fixed. The axes of rotation are an elevation axis which is always horizontal and an azimuth axis which is always vertical. These axes intersect at the centre of the aperture plane of the antenna. The elevation angle of the antenna can be manually adjusted, measured with a protractor and fixed with a locking nut. Rotation of the antenna about the azimuth axis is motorized. The azimuth angle of the antenna at any instance is determined by the resistance of a potentiometer whose shaft is rotated by the antenna. Therefore, the far field amplitude pattern of the antenna can be measured, along what is here called a *constant elevation cut*, by fixing the elevation angle of the antenna, rotating the antenna about its azimuth axis and recording the amplitude of the acoustic pressure at the microphone as a function of the azimuth angle of the antenna. The speed of rotation about the azimuth axis is slow enough for the antenna to transmit about 400 tone bursts in the time taken for the antenna to rotate through 180°.

The measurement microphone is a half inch diameter condenser microphone (Brüel & Kjær Type 4134) and is coupled to a audio frequency spectrometer (Brüel & Kjær Type 2112). The electrical signal from the microphone is a measure of the acoustic pressure at the microphone's head [Rossi, 1988, Sec. 8.1.2]. This electrical signal is amplified and filtered by the spectrometer. The centre frequency of the filter is 12.5 kHz and its bandwidth is 3 kHz. The purpose of the filter is to reject high and low frequency

noise. An important consequence of using the filter is that the measurement apparatus takes a significant time to respond to a tone burst. On starting to receive a nominally constant amplitude tone burst from the microphone, the filtered tone burst only reaches an approximately constant amplitude after about 9 periods of the reference signal. The amplitude of the signal must therefore be determined from the remaining 7 cycles of the tone burst. The signal from the filter is here referred to as the filtered microphone signal.

The amplitude meter generates a signal which is proportional to the amplitude of the most recent tone burst received by the microphone. It consists of an amplitude detector followed by a sample and hold circuit. The amplitude detector consists of a precision AC to DC converter built to a standard design [National, 1980, p. LB8-2]. It outputs a signal which continuously indicates the amplitude of the filtered microphone signal. This amplitude signal is sampled 11.0 ms after the start of each tone burst generated by the waveform generator (Sec. 5.2.1). This time delay allows both for the time taken by the acoustic wave to propagate from the antenna to the microphone and for the output signal from the filter to respond to the signal from the microphone. The sampling is performed by a sample and hold circuit [Millman, 1979, Sec. 16-9], implying that the output signal from the amplitude meter assumes a constant level which is updated to a new constant level whenever the signal from the amplitude detector is sampled.

The pen plotter is fed from the amplitude meter and by a signal whose voltage is proportional to the resistance of the azimuth angle sensing potentiometer. The azimuthal motor drive rotates the antenna by 180° during the measurement of each pattern cut. Therefore, the scale for the axis of the plot corresponding to the azimuth angle is automatically set because the total extent of the plot is constrained to be 180°. It is unnecessary to establish the absolute scale of the amplitude axis because the amplitude data to which the modified Gerchberg-Saxton algorithm is applied are normalized to have a peak value of unity. The zero position of the amplitude scale is established by noting the level of the plot when the amplitude meter is turned off.

A graph, produced by the pen plotter, represents the measured amplitude pattern along a single constant elevation cut. In order to make enough measurements of a far field amplitude pattern to adequately characterize it, the pattern must be measured along many different constant elevation cuts.

5.2.3 Measurement software

The purpose of the measurement software is to convert the data graphed by the pen plotter into a form suitable for being operated upon by the modified Gerchberg-Saxton algorithm. As explained in Chapter 4 and Section 5.1, the modified Gerchberg-Saxton algorithm operates on samples, positioned on a square grid in the u, v plane, of the measured weighted far field amplitude pattern $A_{\mathrm{m}}(u, v)$. These samples can be thought of as being measured samples of the actual weighted far field amplitude pattern $|F_{\mathrm{a}}(u, v)|$. However, in the experiment reported here, such samples of $|F_{\mathrm{a}}(u, v)|$ are not measured directly for two reasons: It is the far field amplitude pattern $|\acute{p}(\theta_{\mathrm{az}}; \theta_{\mathrm{el}})|$, not $|F_{\mathrm{a}}(u, v)|$, that is measured, where θ_{az} and θ_{el} are azimuth and elevation angles respectively; Samples of $|\acute{p}(\theta_{\mathrm{az}}; \theta_{\mathrm{el}})|$ are measured along constant elevation cuts and not on a square grid in the u, v plane. In this section, the measured samples of $|\acute{p}(\theta_{\mathrm{az}}, \theta_{\mathrm{el}})|$, denoted by $|\acute{p}(\theta_{\mathrm{az}}; \theta_{\mathrm{el}})|_{\mathrm{m}}$, are said to be positioned on the *initial sample points.* The sample points on the above-mentioned square grid in the u, v plane are called the *required sample points.* Thus, the software performs three tasks. Firstly each initial sample

point is transformed from the angular coordinates $(\theta_{az};\theta_{el})$ to Cartesian coordinates (u,v). Secondly, the samples $|\acute{p}(u,v)|_m$ are weighted to generate what are here called *initial samples* of $A_m(u,v)$, which are positioned on the initial sample points. Thirdly, samples of $A_m(u,v)$ at the required sample points are estimated from the initial samples of $A_m(u,v)$.

As intimated in the previous paragraph, the angular orientation of the antenna is here denoted by $(\theta_{az};\theta_{el})$ where θ_{az} and θ_{el} are, respectively, the azimuth and elevation angles of the antenna. The orientation $(\theta_{az};\theta_{el}) = (0;0)$ is taken to be that at which the normal to the aperture plane is in the direction of the straight line from the centre of the aperture to the centre of the measuring microphone. This direction is horizontal. The azimuth angle is defined to increase as the antenna is rotated anticlockwise when looked at from above. The elevation angle is defined to increase as the antenna is tilted downwards. Each cut measured by the measurement hardware has a constant elevation in the range $-90°$ to $90°$. The azimuth angle varies from $-90°$ to $90°$ along each cut.

The relationship between an antenna orientation $(\theta_{az};\theta_{el})$ and the corresponding position (u,v) in the far field is expressed as

$$\begin{aligned} u &= \frac{\sin\theta_{az}}{\lambda} \\ v &= \frac{\sin(\theta_{el})\cos(\theta_{az})}{\lambda} \end{aligned} \tag{5.15}$$

Figure 5.4 illustrates this relationship by plotting on the u,v plane the curves corresponding to several constant elevations cuts.

The definitions of u and v in (5.15) imply that, when the centre of the main beam of the antenna is directed towards the microphone, the x and y axes in the aperture plane are, respectively, horizontal and vertical. The values of x and y increase, respectively, to the right and up the aperture plane, when viewed from the microphone. The origin of the aperture plane coincides with the intersection of the azimuth and elevation axes. The centre speaker of the antenna is centred on the origin of the aperture plane.

Before the software can manipulate the data, they must be input into the computer. This is achieved with the aid of a digitizer unit (GTCO corporation, Micro DIGI-PAD) consisting of a tablet and crosshair. Each plot from the pen plotter is placed in turn on the tablet. By moving the crosshair along the plotted curve by hand, the coordinates of between 100 and 150 points on the curve are automatically passed to the computer by the digitizer unit. All relevant information about the azimuth angle and the relative amplitude of the samples positioned along each constant elevation cut are transferred in this way to computer memory. The elevation angle for each cut is also fed to the computer.

The set of samples obtained from the digitizer unit constitute $|\acute{p}(\theta_{az};\theta_{el})|_m$. By invoking (5.15), each initial sample point can be transformed from $(\theta_{az};\theta_{el})$ into (u,v) so that $|\acute{p}(u,v)|_m = |\acute{p}(\theta_{az};\theta_{el})|_m$. It follows from (5.14) that, at each initial sample point,

$$A_m(u,v) = \frac{|\acute{p}(u,v)|_m}{\lambda w(u,v)} \tag{5.16}$$

where $w(u,v)$ is defined in (2.28).

It is intimated in (3.19) that the required sample points are the points (u,v) for which

$$u = m\Delta_u \quad \text{and} \quad v = n\Delta_v \tag{5.17}$$

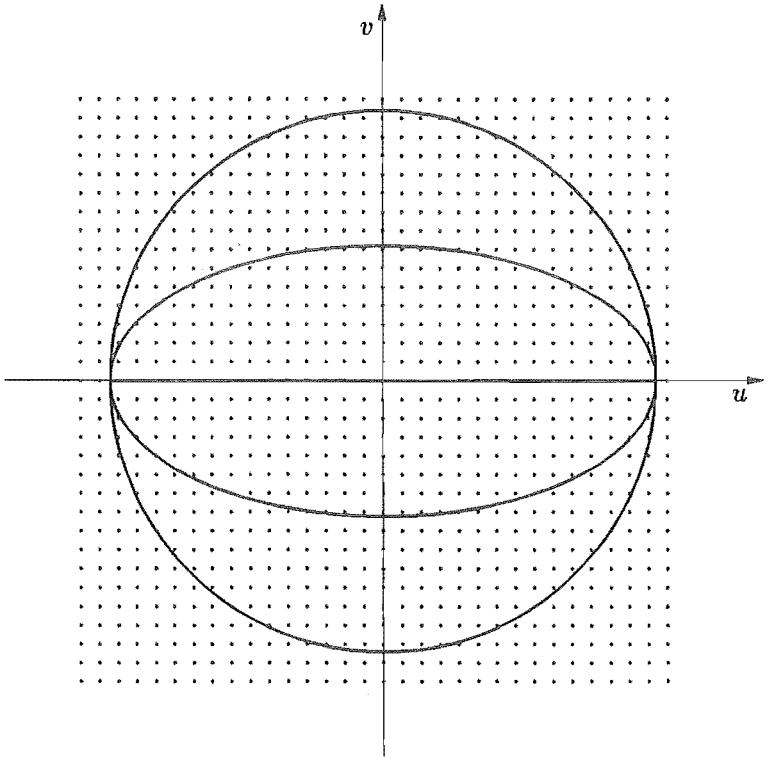


Figure 5.4 Curves, in the u, v plane, corresponding to constant elevation cuts for which $\theta_{el} = -90°, -30°, 0°, 30°$ and $90°$. The measurement hardware plots the variation of the far field amplitude pattern along such curves. The dots represent points on a sampling grid. The purpose of the measurement software is to calculate the weighted far field amplitude at each of these points.

where Δ_u and Δ_v are the spacings of the samples in the u and v directions respectively. The integers m and n each range from $-M/2$ to $M/2 - 1$ and $-N/2$ to $N/2 - 1$, respectively, where M and N are the numbers of samples in the u and v directions respectively.

As depicted in Figure 5.4, the constant elevation cuts describe semi-ellipses in the u, v plane as θ_{az} varies from $-90°$ to $90°$. Consider a line, parallel to the v axis, which cuts the u axis at $u = m\Delta_u$. Provided $u < (1/\lambda)$, this line intersects all of the constant elevation cuts. Consider the intersection of the line with the cut at elevation θ_{el}. The intersection is at the point $(m\Delta_v, v)$ where, by substituting the first equation of (5.15) into the second equation of (5.15),

$$v = \sin(\theta_{el}) \frac{[1 - (\lambda m \Delta_u)^2]^{1/2}}{\lambda} \tag{5.18}$$

The value of $A_m(m\Delta_u, v)$ is calculated by linearly interpolating the initial samples of $A_m(u, v)$ along the cut. This approach can be invoked to calculate the values of $A_m(m\Delta_u, v)$ for all m on all the constant elevation cuts. Linear interpolation is employed, instead of some higher order type of interpolation, because it is simple and is sufficiently accurate on account of the initial sample points being spaced so closely along each cut.

The elevation angles of the cuts are chosen such that, from cut to cut, the value of $\sin(\theta_{\mathrm{el}})$ changes by equal increments. This implies, from (5.18), that the values of $A_{\mathrm{m}}(m\Delta_u, v)$ can be inferred, from the initial samples of $A_{\mathrm{m}}(u, v)$, at equally spaced increments in v along the line $u = m\Delta_u$. The value of $A_{\mathrm{m}}(m\Delta_u, n\Delta_v)$ is obtained for each n by sinc interpolating the square of $A_{\mathrm{m}}(m\Delta_u, v)$. The justification for doing this is now outlined. Neglecting measurement inaccuracies, $A_{\mathrm{m}}(u, v) = |F_{\mathrm{a}}(u, v)|$, implying from (3.42) that the inverse Fourier transform of $[A_{\mathrm{m}}(u, v)]^2$ is the autocorrelation $ff_{\mathrm{a}}(x, y)$ of the actual aperture field distribution $f_{\mathrm{a}}(x, y)$. Assuming that $f_{\mathrm{a}}(x, y)$ is compact (Sec. 3.4.1.1) then, from (3.43), $ff_{\mathrm{a}}(x, y)$ is also compact. From the definition of the inverse Fourier transform operator (Table 3.3), the one-dimensional inverse Fourier transform of $[A_{\mathrm{m}}(m\Delta_u, v)]^2$ is

$$\int [A_{\mathrm{m}}(m\Delta_u, v)]^2 e^{-j2\pi vy}\, dy = \int ff_{\mathrm{a}}(x, y) e^{-j2\pi m\Delta_u x}\, dx \tag{5.19}$$

for each integer m. The right side of (5.19) must be compact since $ff_{\mathrm{a}}(x, y)$ is compact. It follows from the discussion in Section 3.4.1.3 that the value of $[A_{\mathrm{m}}(m\Delta_u, v)]^2$ can be determined for any v by sinc interpolating (cf. (3.23)) values of $[A_{\mathrm{m}}(m\Delta_u, v)]^2$ specified at equally spaced values of v. However, there is the proviso (Sec. 3.4.1.3) that the increments in v must be no greater than $1/(2L_y^{f_{\mathrm{a}}})$, where $L_y^{f_{\mathrm{a}}}$ is the extent of $f_{\mathrm{a}}(x, y)$ in the y direction. It is apparent from (5.18) that this is ensured if the increment in $\sin(\theta_{\mathrm{el}})$ between successive constant elevation cuts is less than or equal to $\lambda/(2L_y^{f_{\mathrm{a}}})$.

The algorithm detailed above generates values of $A_{\mathrm{m}}(u, v)$ at all the required sample points $(u, v) = (m\Delta_u, n\Delta_v)$ for which $(u^2 + v^2) < (1/\lambda)^2$. The algorithm sets to zero the values of $A_{\mathrm{m}}(u, v)$ at all the required sample points for which $(u^2 + v^2) \geq (1/\lambda)^2$ (cf. Sec. 2.1.3.2).

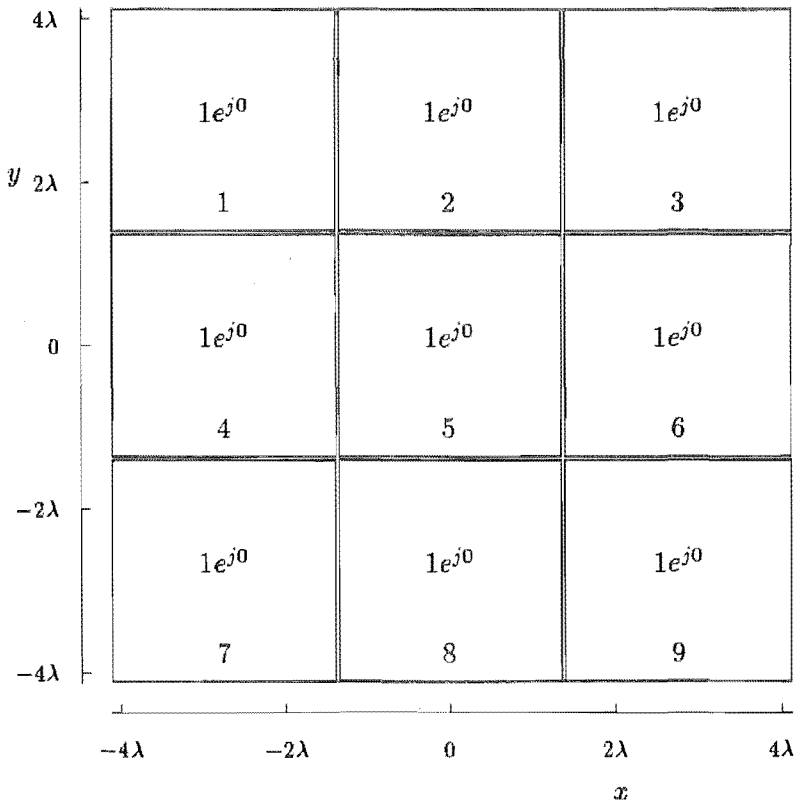
5.3 RESULTS

Measurements have been made of the far field amplitude patterns corresponding to two different configurations of the acoustic antenna. These configurations and various details of the measurements are described in Section 5.3.1. Section 5.3.2 discusses some processing of the far field data and why it is necessary. A way of determining the extent of the aperture field distribution is also presented. Methods for evaluating the results generated by the modified Gerchberg-Saxton algorithm are described in Section 5.3.3. Section 5.3.4 presents the results of processing the measured data in the way described in Section 5.3.2. Section 5.3.5 presents the results generated by the modified Gerchberg-Saxton algorithm when it is applied to the processed measured data.

5.3.1 Details of two measurements

Two configurations of the antenna are considered here. They are denoted configuration A and configuration B and are illustrated in Figure 5.5. Recall from Section 5.2.1 that the configuration of the antenna is determined by the amplitudes and the phases of the signals fed to the nine speakers comprising the antenna. In configuration A, the same amplitudes and phases are fed to all the speakers. Configuration B is the same as configuration A with the exception that the phase of the signal feeding speaker 1 (i.e. the top left one — see Fig. 5.5(b)) is delayed by $\pi/2$ rad. For the remainder of this chapter, the superscripts and subscripts 'A' and 'B' refer to configurations A and B respectively.

(a)



(b)

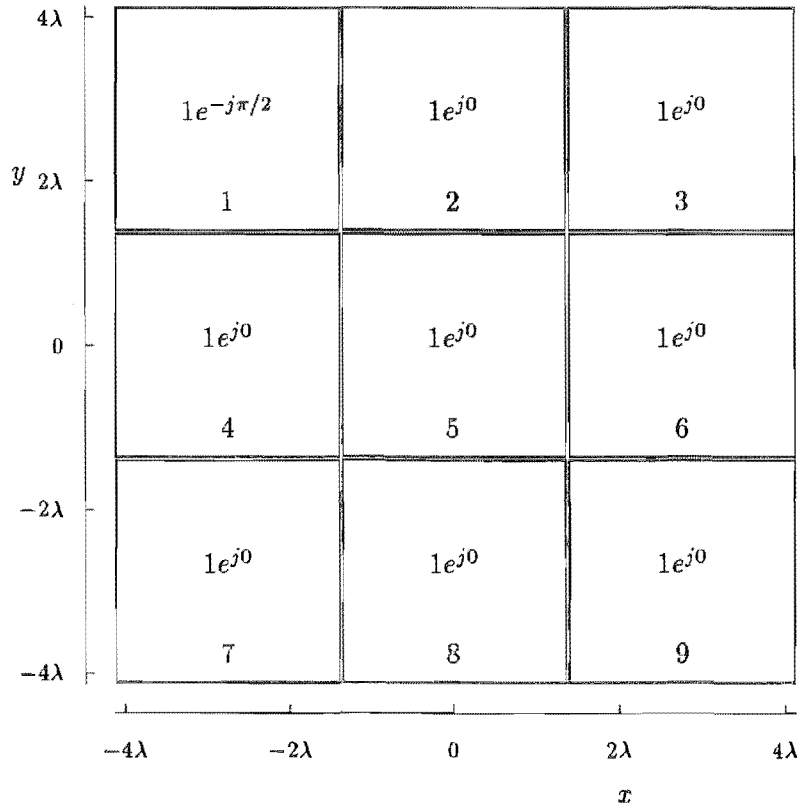


Figure 5.5 Two antenna configurations: (a) configuration A; (b) configuration B. The squares indicate the positions of the speakers comprising the antenna. The complex numbers indicate the relative amplitudes and phases of the signals feeding the speakers. The integer numbers identify the speakers.

In this section, as well as in Sections 5.3.2 and 5.3.5, the units on the u and v axes are each chosen to be $1/\lambda$. Correspondingly, the units on the x and y axes are each λ.

The physical extent of the aperture in each of the x and y directions is 8.27λ (see Fig. 5.5). Therefore, the value of L^{f_a} has to be assumed to be 8.27λ unless, and until, a more appropriate value for the extent of $f_a(x, y)$ can be determined. It transpires, in fact, that once the far field amplitude pattern has been measured, an improved estimate of L^{f_a} can be generated (as explained in Sec. 5.3.2).

For each configuration, $|\acute{p}(\theta_{az}; \theta_{el})|$ was measured along 35 constant elevation cuts, as defined in Section 5.2.3. In order to fulfil the requirements of the measurement software (Sec. 5.2.3), the increment in $\sin\theta_{el}$ between successive cuts was set to $\lambda/(2L^{f_a})$.

To check the accuracy of the measurements, $|\acute{p}(\theta_{az}; \theta_{el})|$ was measured along the cut for which $\theta_{el} = 0°$ before, and again after, it was measured along the remaining 34 constant elevations cuts. The rms difference between the first $|\acute{p}(\theta_{az}; 0)|_m$ and the second $|\acute{p}(\theta_{az}; 0)|_m$ was about 0.015 times the peak value of $|\acute{p}(\theta_{az}; \theta_{el})|_m$. Because θ_{el} was measured to an accuracy of $\pm 0.125°$, it is estimated that this introduced an rms error of about 0.005 times the peak value of $|\acute{p}(\theta_{az}; \theta_{el})|_m$. The error introduced during the operation of the digitizer unit (Sec. 5.2.3) is taken to be the line width of the plot produced by the pen plotter. This error is about 0.005 times the peak value of $|\acute{p}(\theta_{az}; \theta_{el})|_m$. Overall, it is therefore estimated that $A_m^A(u, v)$ and $A_m^B(u, v)$ are accurate to within an rms value of 0.025 times the peak values of their respective main beams.

Equations (3.37) and (3.44) specify conditions for the size and sample spacing of the far field and aperture plane sampling grids. Recall that the aperture plane sampling grid is defined by the sample spacings Δ_x and Δ_y and by the numbers of samples M and N in the x and y directions respectively. Similarly, the far field sampling grid is defined by Δ_u, Δ_v, M and N. The particular implementation of the FFT algorithm, invoked to perform any Fourier transform operations (Sec. 3.4.1.4), requires that both M and N be powers of 2.

The data to which the modified Gerchberg-Saxton algorithm is applied (Sec. 5.3.5) are sampled over the aperture and far field sampling grids defined by

$$\Delta_x = \Delta_y = 0.55\lambda, \quad \Delta_u = \Delta_v = 0.057/\lambda \quad \text{and} \quad M = N = 32 \tag{5.20}$$

The aperture sampling grid is such that each speaker is centred on a sample point. The speaker centres are 5 sample spacings apart. It follows from (3.22) and (5.20) that $A_m^A(u, v)$ and $A_m^B(u, v)$ are oversampled by factors of 2.13 in both the u and v directions.

5.3.2 Methods for processing far field amplitude data

This section is concerned with the amplitude data $A_m(u, v)$ generated by the measurement hardware and software. It is explained that some of the data comprising $A_m(u, v)$ should be removed because it does not relate to what the modified Gerchberg-Saxton algorithm requires as input. It is suggested that the data be removed by truncating $A_m(u, v)$. This section finishes by proposing a method for determining the value of L^{f_a} from the truncated $A_m(u, v)$.

Let θ be the angle between the positive z axis and the direction corresponding to any particular pair of values of u ad v. It follows from (2.28) that, as θ increases to $90°$, $(u^2 + v^2)^{1/2}$ increases to $1/\lambda$ and $w(u, v)$ decreases to 0. The presence of $w(u, v)$ in (5.13) implies that, according to the theory invoked to derive (5.13), $\acute{p}(u, v)$ must be expected to tend to zero as θ approaches $90°$. It turns out, however, that the measured $|\acute{p}(u, v)|$ is actually non-zero for $\theta = 90°$, implying that a finite acoustic field is radiated

in these directions. An important consequence of this, which follows from (5.16), is that, unless some appropriate precaution is taken, $A_{\mathrm{m}}(u,v)$ would become infinitely large as θ tends to 90°.

The modified Gerchberg-Saxton algorithm requires $A_{\mathrm{m}}(u,v)$ to be an approximation to $|F_{\mathrm{a}}(u,v)|$ where $F_{\mathrm{a}}(u,v) = \mathrm{FT}\{f_{\mathrm{a}}(x,y)\}$ and where $f_{\mathrm{a}}(x,y)$ is compact. Unfortunately, the discussion in the previous paragraph implies that $A_{\mathrm{m}}(u,v)$ is a poor approximation to $|F_{\mathrm{a}}(u,v)|$ in directions $\theta \approx 90°$. However, there is no reason to suggest that $A_{\mathrm{m}}(u,v)$ is not an adequate approximation to $|F_{\mathrm{a}}(u,v)|$ in directions for which θ is appreciably different from 90°. In fact, in these directions, $A_{\mathrm{m}}(u,v)$ tends to be typical of the Fourier transform amplitude of a compact aperture field distribution. This implies that there must be fields in the aperture, other than $f_{\mathrm{a}}(x,y)$, which radiate predominantly in directions $\theta \approx 90°$. Because it is $f_{\mathrm{a}}(x,y)$, not these other aperture fields, which is of primary concern in this thesis, it seems reasonable to find a way of removing the measured data corresponding to $\theta \approx 90°$. One way of achieving this is to truncate $A_{\mathrm{m}}(u,v)$ in the region of the u,v plane lying outside a disk, denoted by $S^{A_{\mathrm{m}}}$, centred on the origin and of diameter $D^{A_{\mathrm{m}}} < 2/\lambda$. The question of what value should be chosen for $D^{A_{\mathrm{m}}}$ is addressed in the next paragraph.

It is intimated in Section 4.7.1 that the smaller the differences between $A_{\mathrm{m}}(u,v)$ and $|F_{\mathrm{a}}(u,v)|$, the more compact is $ff_{\mathrm{m}}(x,y)$ (defined in (4.48)). The following procedure identifies that truncated $A_{\mathrm{m}}(u,v)$ for which $ff_{\mathrm{m}}(x,y)$ is most compact. Many different values for $D^{A_{\mathrm{m}}}$ are considered in turn. For each value of $D^{A_{\mathrm{m}}}$, $A_{\mathrm{m}}(u,v)$ is truncated outside $S^{A_{\mathrm{m}}}$. The truncated data are then squared and inverse Fourier transformed to form $ff_{\mathrm{m}}(x,y)$ (cf. (4.48)). Recall from Section 4.7.1 that the autocorrelation error $\mathcal{E}^{\mathrm{auto}}$ provides a measure of the compactness of $ff_{\mathrm{m}}(x,y)$. In terms which are relevant to this discussion, the autocorrelation error $\mathcal{E}^{\mathrm{auto}}$ is here defined to be (cf. (4.50))

$$\mathcal{E}^{\mathrm{auto}} = \frac{1}{ff_{\mathrm{m}}(0,0)} \left[\frac{\iint_{(x,y)\notin S^{\mathrm{auto}}} |ff_{\mathrm{m}}(x,y)|^2 \, dx \, dy}{\iint_{(x,y)\notin S^{\mathrm{auto}}} dx \, dy} \right]^{1/2} \tag{5.21}$$

where S^{auto} is taken to be the square region of the x,y plane, the length of whose sides is $2L^{f}$. Ideally L^{f} would be identical to $L^{f_{\mathrm{a}}}$, but, as explained in Section 5.3.1, the value of L^{f} has to be here taken to be 8.27λ. The value for $D^{A_{\mathrm{m}}}$ is chosen to be that which corresponds to the smallest value of $\mathcal{E}^{\mathrm{auto}}$. In the next paragraph it is assumed that $A_{\mathrm{m}}(u,v)$ has been truncated according to this chosen value of $D^{A_{\mathrm{m}}}$.

It is now explained how the value of $L^{f_{\mathrm{a}}}$ can be estimated from $A_{\mathrm{m}}(u,v)$. Note that, because S^{auto} is defined above in terms of L^{f}, $\mathcal{E}^{\mathrm{auto}}$ can be computed for many different values of L^{f}. Recall that $ff_{\mathrm{a}}(x,y)$ is the autocorrelation of $f_{\mathrm{a}}(x,y)$. It is assumed here to be exactly compact so that it vanishes outside S^{auto} when $L^{f} = L^{f_{\mathrm{a}}}$. It is also assumed that $[ff_{\mathrm{m}}(x,y) - ff_{\mathrm{a}}(x,y)]$ is a noise-like distribution with an rms value of, say, $\xi ff_{\mathrm{m}}(0,0)$. It follows from (5.21) that $\mathcal{E}^{\mathrm{auto}} \approx \xi$ for all $L^{f} \geq L^{f_{\mathrm{a}}}$. However, when $L^{f} < L^{f_{\mathrm{a}}}$, $\mathcal{E}^{\mathrm{auto}} > \xi$. This suggests that $L^{f_{\mathrm{a}}}$ should be estimated as follows. Compute $\mathcal{E}^{\mathrm{auto}}$ for many different values of L^{f}. Then set $L^{f_{\mathrm{a}}}$ to be the smallest value for which $\mathcal{E}^{\mathrm{auto}}$ is approximately constant for all $L^{f} \geq L^{f_{\mathrm{a}}}$.

5.3.3 Methods for evaluating results

In Section 5.3.5 the modified Gerchberg-Saxton algorithm is applied to the experimentally obtained far field data $A^{\mathrm{A}}_{\mathrm{m}}(u,v)$ and $A^{\mathrm{B}}_{\mathrm{m}}(u,v)$. Recall that the purpose of the

modified Gerchberg-Saxton algorithm is to generate estimates, phase$\{f_e^A(x,y)\}$ and phase$\{f_e^B(x,y)\}$ respectively, of the phase of the image-forms of $f_a^A(x,y)$ and $f_a^B(x,y)$. In order to evaluate the modified Gerchberg-Saxton algorithm, the accuracy of these estimates must be ascertained. Unlike in computer simulations (Chap. 4), the aperture phase error $\mathcal{E}^{ap}$ defined in (4.21) cannot be computed here, because phase$\{f_a(x,y)\}$ is not available. However, for the particular experiments reported here, the relative phases of the signals feeding each speaker in the antenna are known, as are the relative positions of the speakers. This provides limited information about phase$\{f_a^A(x,y)\}$ and phase$\{f_a^B(x,y)\}$ with which to compare phase$\{f_e^A(x,y)\}$ and phase$\{f_e^B(x,y)\}$. This section states all the information about $f_a^A(x,y)$ and $f_a^B(x,y)$ which is known before $f_e^A(x,y)$ and $f_e^B(x,y)$ are generated. This information can be used to estimate the accuracy of phase$\{f_e^A(x,y)\}$ and phase$\{f_e^B(x,y)\}$ once they have been generated. Two methods for doing this are described in this section. For convenience it is assumed that $f_e^A(x,y)$ and $f_e^B(x,y)$, rather than $\tilde{f}_e^A(x,y)$ and $\tilde{f}_e^B(x,y)$, are the estimates of $f_a^A(x,y)$ and $f_a^B(x,y)$ respectively.

Consider first the situation in which the signals feeding all but the l^{th} speaker of the antenna are turned off. The resulting aperture field distribution is here called the *speaker aperture field distribution* of that speaker and is denoted by $f_{sl}(x,y)$. Now consider the situation in which the amplitudes of the signals feeding the nine speakers are equal to each other, but the phases of these signals are arbitrary. Assuming that mutual coupling between the speakers is negligible, the aperture field distribution $f_a(x,y)$ of the antenna is

$$f_a(x,y) = \sum_{l=1}^{9} f_{sl}(x,y)e^{j\psi_{sl}} \tag{5.22}$$

where ψ_{sl} is the relative phase of the signal feeding the l^{th} speaker and where $f_{s5}(x,y)$ is the speaker aperture field distribution of the central speaker. Assuming that the electromechanical properties of the speakers are also identical, it follows that

$$f_{sl}(x,y) = f_{s5}(x - x_{sl}, y - y_{sl})e^{j\psi_{sl}} \quad \text{for } l = 1 \text{ to } 9 \tag{5.23}$$

where (x_{sl}, y_{sl}) is the position of the centre of the l^{th} speaker.

The physical symmetry of the speakers implies that $f_{s5}(x,y)$ is circularly symmetric, by which it is meant that $\partial f(r\cos\phi, r\sin\phi)/\partial\phi \equiv 0$. For reasons given in Section 3.4.1.1 it is assumed that $f_{s5}(x,y)$ is approximately compact with an extent $L^{f_{s5}}$ in both the x and y directions. It follows from (5.23) and (5.22) that L^{f_a} is related to $L^{f_{s5}}$ by

$$L^{f_a} = x_{s6} - x_{s4} + L^{f_{s5}} \tag{5.24}$$

Note that $x_{s6} = -x_{s4} = 2.76\lambda$. It turns out (see Sec. 5.3.5) that L^{f_a} is small enough to imply, from (5.24), that $L^{f_{s5}} < x_{s6}$. The implication of this is that the different speaker aperture field distributions do not, in fact, overlap.

The first method for determining the accuracy to which phase$\{f_e^A(x,y)\}$ and phase$\{f_e^B(x,y)\}$ approximate phase$\{f_a^A(x,y)\}$ and phase$\{f_a^B(x,y)\}$, respectively, is based on knowledge of the differences between configurations A and B. It need not be assumed here that the speakers behave identically (i.e. (5.23) need not hold), but it is assumed that the extent of $f_{sl}(x,y)$ for each l is less than x_{sl}. From (5.22) and the descriptions in Section 5.3.1 of configurations A and B, the actual aperture field phase distributions corresponding to these two configurations are

$$\begin{aligned} \text{phase}\{f_a^A(x,y)\} &= \sum_{l=1}^{9} \text{phase}\{f_{sl}(x,y)\} \\ \text{phase}\{f_a^B(x,y)\} &= \text{phase}\{f_{s1}(x,y)e^{-j\pi/2}\} + \sum_{l=2}^{9} \text{phase}\{f_{sl}(x,y)\} \end{aligned} \tag{5.25}$$

Therefore, the phase difference $\Delta\psi^{B-A}(x,y)$ between $f_a^A(x,y)$ and $f_a^B(x,y)$ is

$$\Delta\psi^{B-A}(x,y) = \begin{cases} -j\pi/2 & \text{where } [(x-x_{s1})^2 + (y-y_{s1})^2]^{1/2} \leq x_{s6} \\ 0 & \text{elsewhere} \end{cases} \tag{5.26}$$

An estimate $\Delta\psi_e^{B-A}(x,y)$ of $\Delta\psi^{B-A}(x,y)$ can be calculated from $f_e^A(x,y)$ and $f_e^B(x,y)$, generated by the modified Gerchberg-Saxton algorithm, as follows:

$$\Delta\psi_e^{B-A}(x,y) = \text{phase}\{f_e^B(x,y)\} - \text{phase}\{f_e^A(x,y)\} \tag{5.27}$$

Therefore, the rms difference between $\Delta\psi_e^{B-A}(x,y)$ and $\Delta\psi^{B-A}(x,y)$ provides one indication of the accuracy of phase$\{f_e^A(x,y)\}$ and phase$\{f_e^B(x,y)\}$.

There is an alternative method for determining the accuracy of phase$\{f_e^A(x,y)\}$ and phase$\{f_e^B(x,y)\}$, which has the advantage that it treats the two configurations independently. However, it relies on the assumption that the speakers behave identically (i.e. (5.23) is assumed to hold). Substituting (5.23) into (5.25) yields

$$\begin{aligned} \text{phase}\{f_a^A(x,y)\} &= \sum_{l=1}^{9} \text{phase}\{f_{s5}(x-x_{sl}, y-y_{sl})\} \\ \text{phase}\{f_a^B(x,y)\} &= \text{phase}\{f_{s5}(x-x_{s1}, y-y_{s1})\}e^{-j\pi/2} \\ &\quad + \sum_{l=2}^{9} \text{phase}\{f_{s5}(x-x_{sl}, y-y_{sl})\} \end{aligned} \tag{5.28}$$

Even if phase$\{f_{s5}(x,y)\}$ is not known, phase$\{f_e^A(x,y)\}$ and phase$\{f_e^B(x,y)\}$ should respectively have the forms of the first and second equations of (5.28). Note, in particular, that the samples of phase$\{f_e^A(x,y)\}$ positioned at, say, the centres of each of the speakers should be equal to each other. Similarly, the samples of [phase$\{f_e^B(x,y)\}$ − $\Delta\psi^{B-A}(x,y)$] positioned at the centres of the speakers can also be expected to be equal to each other, where $\Delta\psi^{B-A}(x,y)$ is defined in (5.26). The rms variations of these two sets of sample values provide indications of the accuracy of phase$\{f_e^A(x,y)\}$ and phase$\{f_e^B(x,y)\}$ respectively.

5.3.4 Processing the far field data

In this section it is shown how the data $A_{\mathrm{m}}^{\mathrm{A}}(u,v)$ and $A_{\mathrm{m}}^{\mathrm{B}}(u,v)$ are processed. Recall that $A_{\mathrm{m}}^{\mathrm{A}}(u,v)$ and $A_{\mathrm{m}}^{\mathrm{B}}(u,v)$ are obtained from the measurement hardware and software (Secs. 5.2.2, 5.2.3 and 5.3.1). They are processed by invoking the methods developed in Section 5.3.2. It is also indicated in this section how a distribution for $|f_{\mathrm{d}}(x,y)|$ is specified.

Figure 5.6(a) depicts graphs of $\mathcal{E}^{\mathrm{auto}}$ versus 16 different values of $D^{A_{\mathrm{m}}}$ for configurations A and B. For both of these configurations, $\mathcal{E}^{\mathrm{auto}}$ is minimum when $D^{A_{\mathrm{m}}} = 1.77/\lambda$. Accordingly, it is assumed throughout the remainder of this chapter that $A_{\mathrm{m}}^{\mathrm{A}}(u,v)$ and $A_{\mathrm{m}}^{\mathrm{B}}(u,v)$ are truncated (see Sec. 5.3.2) with this value of $D^{A_{\mathrm{m}}}$. Figures 5.6(b) and (c) show cuts through (the truncated) $A_{\mathrm{m}}^{\mathrm{A}}(u,v)$ and $A_{\mathrm{m}}^{\mathrm{B}}(u,v)$ respectively.

Figure 5.7(a) depicts graphs of $\mathcal{E}^{\mathrm{auto}}$ against 9 different values of L^f, for configurations A and B. Both graphs are relatively steep for $L^f < 7.17\lambda$ and are relatively level for $L^f > 7.17\lambda$. Therefore, following the reasoning given in Section 5.3.2, it is estimated that $L^{f_{\mathrm{a}}} = 7.17\lambda$. This estimate is seen to be justified by inspecting Figures 5.7(b) and (c), which show that $ff_{\mathrm{m}}^{\mathrm{A}}(x,y)$ and $ff_{\mathrm{m}}^{\mathrm{B}}(x,y)$, respectively, are negligible for $|x| > 7.17\lambda$.

Recall from Section 4.1.1 that, when no design information is available, one way of specifying $|f_{\mathrm{d}}(x,y)|$ is to let it be a guess at the distribution $|f_{\mathrm{a}}(x,y)|$. Taking this approach, $|f_{\mathrm{d}}(x,y)|$ is here defined to be (cf. (5.22) and (5.23))

$$|f_{\mathrm{d}}(x,y)| = \sum_{l=1}^{9} |f_{\mathrm{sd}}x - x_{\mathrm{s}l}, y - y_{\mathrm{s}l})| \tag{5.29}$$

where $|f_{\mathrm{sd}}x,y)|$ is the following cone-shaped distribution:

$$f_{\mathrm{sd}}x,y) = \begin{cases} 1 - 2(x^2+y^2)^{1/2}/L^{f_{\mathrm{s5}}} & \text{where } (x^2+y^2)^{1/2} \leq L^{f_{\mathrm{s5}}}/2 \\ 0 & \text{elsewhere} \end{cases} \tag{5.30}$$

and where, from (5.24), $L^{f_{\mathrm{s5}}} = 1.65\lambda$. This distribution for $|f_{\mathrm{d}}(x,y)|$ accords with all of the information about $f_{\mathrm{a}}(x,y)$ mentioned in Section 5.3.3. Note that, because $|f_{\mathrm{d}}(x,y)|$ does not depend on $\psi_{\mathrm{s}l}$, the same $|f_{\mathrm{d}}(x,y)|$ is applicable to both of the configurations A and B. The centre 15 by 15 samples of $|f_{\mathrm{d}}(x,y)|$ are depicted in Figure 5.8. The resolution in the aperture plane is such that $|f_{\mathrm{d}}(x,y)|$ is represented by 9 non-zero samples per speaker.

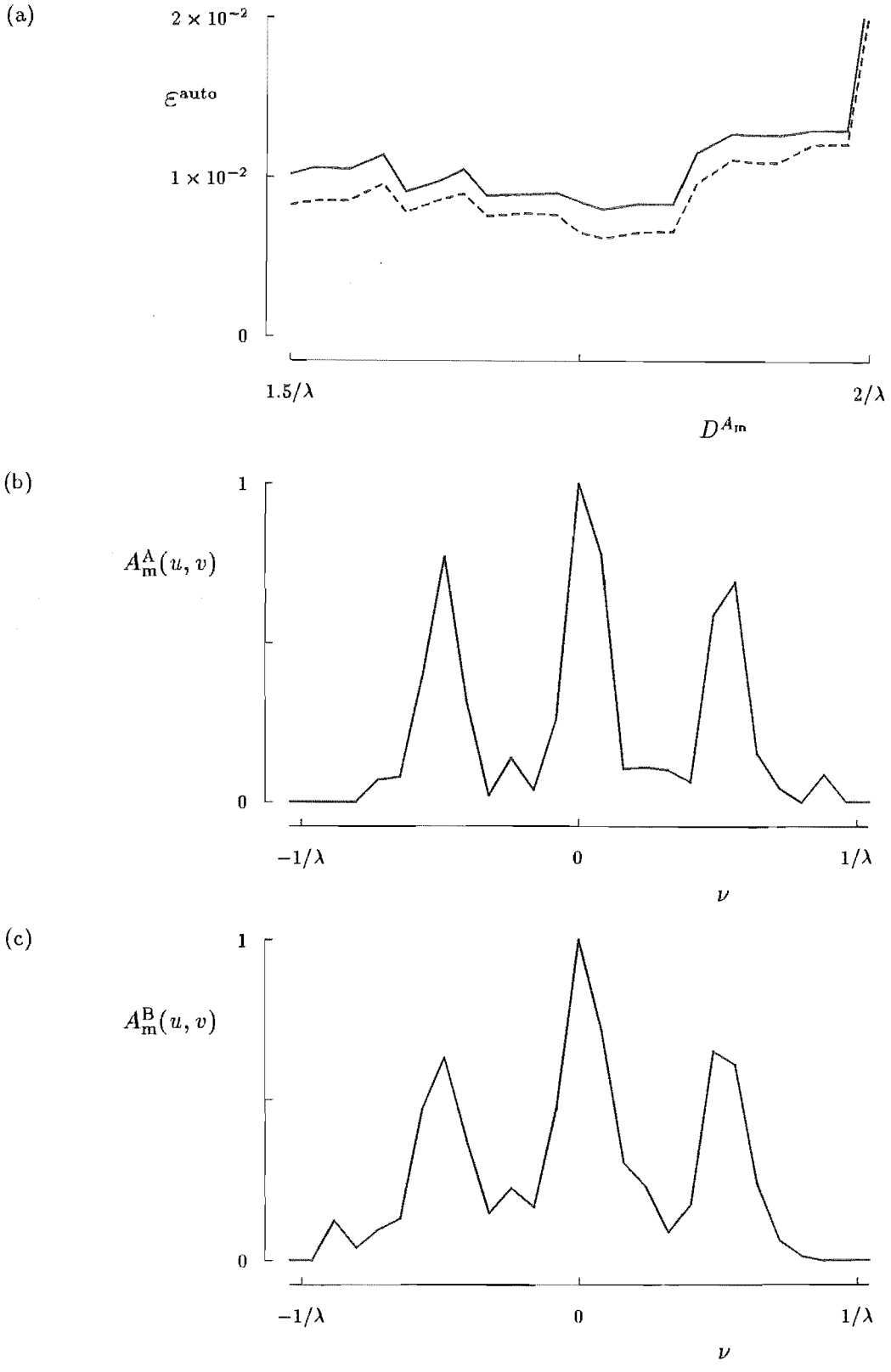


Figure 5.6 Measured weighted far field amplitude distributions: (a) graph of autocorrelation error versus the diameter of a disc in the u, v plane outside of which $A_m^A(u,v)$ (solid curve) and $A_m^B(u,v)$ (dashed curve) are truncated; (b) cut through the truncated $A_m^A(u,v)$ for $D^{A_m} = 1.77/\lambda$; (c) cut through the truncated $A_m^B(u,v)$ for $D^{A_m} = 1.77/\lambda$. $A_m^A(u,v)$ and $A_m^B(u,v)$ are normalized to have maximum values of unity. The variable ν, which identifies position along a diagonal in the far field plane, is defined in (4.65).

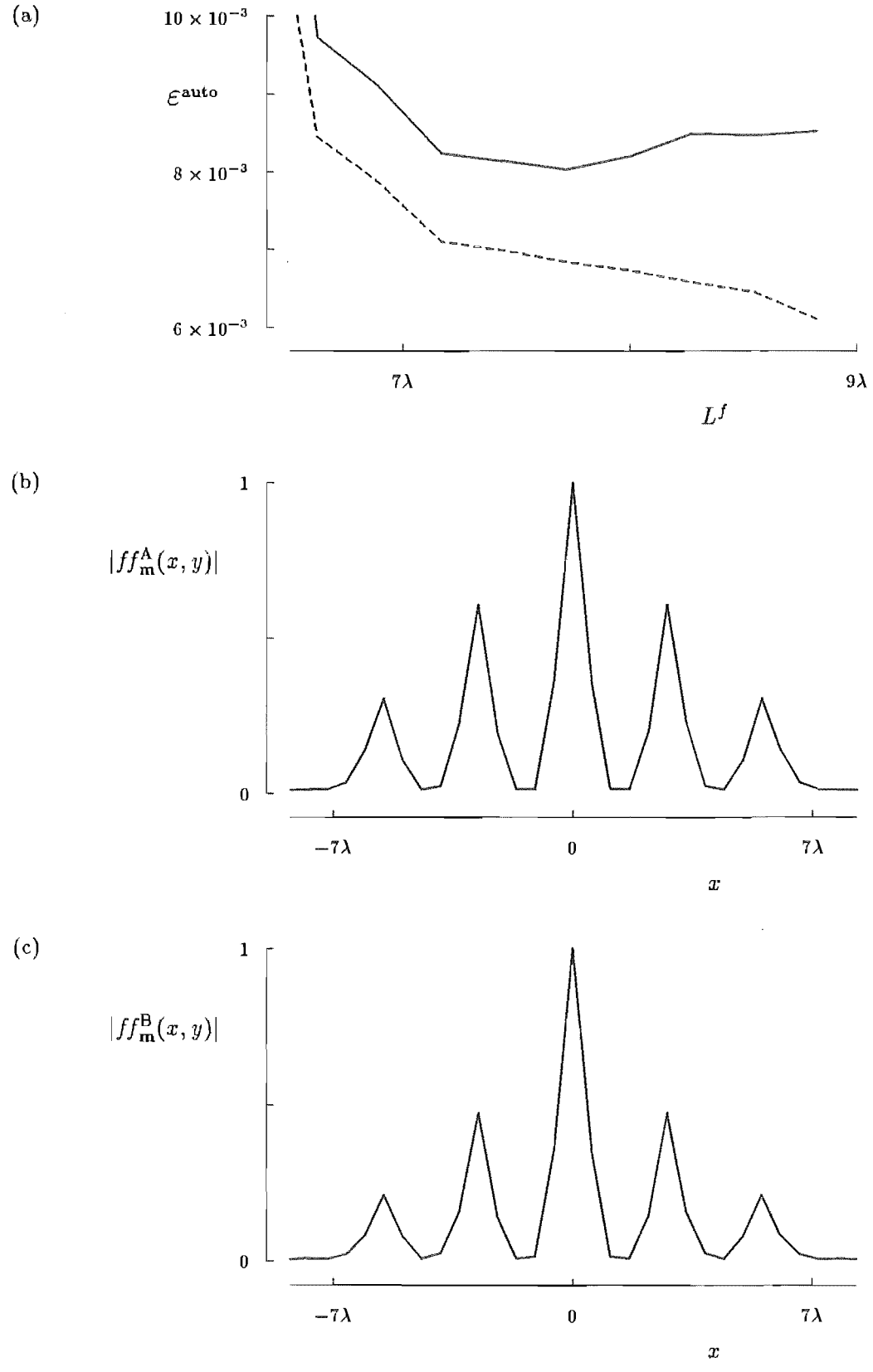


Figure 5.7 Determining the extent of $f_{\text{a}}(x,y)$ from the approximate autocorrelation distributions: (a) graph of autocorrelation error versus L^f (see (5.21)) for configurations A (solid curve) and B (dashed curve); (b) cut through the approximate autocorrelation $ff_{\text{m}}^{\text{A}}(x,y)$; (c) cut through the approximate autocorrelation $ff_{\text{m}}^{\text{B}}(x,y)$.

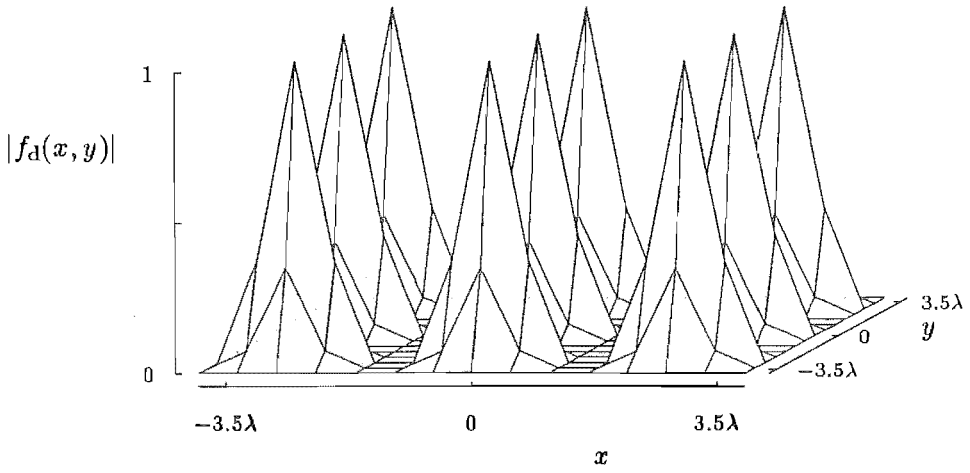


Figure 5.8 Design aperture field amplitude distribution $|f_d(x,y)|$, chosen for both configuration A and configuration B.

5.3.5 Applying the modified Gerchberg-Saxton algorithm

Having obtained measured data $A_m(u,v)$ and design data $|f_d(x,y)|$ (Sec. 5.3.4) for configurations A and B (Sec. 5.3.1), the modified Gerchberg-Saxton algorithm was invoked. In this section the results are illustrated and are evaluated by the methods described in Section 5.3.3.

Because $A_m^A(u,v)$ and $A_m^B(u,v)$ are both truncated (Sec. 5.3.2), the form of the modified Gerchberg-Saxton algorithm to be utilized here is the extrapolating composite algorithm (Sec. 4.7.3.3). Recall that each iteration of the extrapolating composite algorithm is described by either (4.36), (4.38) or (4.31) but is modified by either (4.54) or (4.55). For the runs of the extrapolating composite algorithm described here, $\Gamma_{thres}^f = 0.6A_m^A(0,0)$, S^{A_m} is a disk in the u,v plane of diameter $D^{A_m} = 1.77/\lambda$ and S^{aper} is the region of the x,y plane within a square the length of whose sides are $L^{f_a} = 7.17\lambda$. The extrapolating composite algorithm was run once for configuration A and once for configuration B. The two runs converged to values of the far field error $\mathcal{E}^{fa}$ of 0.025 and 0.024 respectively.

The extrapolating composite algorithm generated $f_e^A(x,y)$ and $f_e^B(x,y)$, which are estimates of the image-forms of $f_a^A(x,y)$ and $f_a^B(x,y)$ respectively. It is not an aim of this experiment to resolve the ambiguities between $f_e^A(x,y)$ and $\tilde{f}_e^A(x,y)$ nor those between $f_e^B(x,y)$ and $\tilde{f}_e^B(x,y)$. The aim is rather to generate, on the basis of the accuracy criteria established in Section 5.3.3, the best estimates of $f_a^A(x,y)$ and $f_a^B(x,y)$. It is here asserted that $\tilde{f}_e^A(x,y)$ and $f_e^B(x,y)$ are these estimates.

Figure 5.9 depicts the phases of $\tilde{f}_e^A(x,y)$ and $f_e^B(x,y)$, as well as their phase difference $\Delta\psi_e^{B-A}(x,y)$. Note that, purely for purposes of illustration, all phase distributions are multiplied by -1 before being plotted in the figures presented in this section. This is done so that it is easy to discern any phase lag, in the speaker aperture field distribution of speaker 1, relative to the phases of the other speaker aperture field distributions. Note also, that a constant has been added to each phase distribution presented in this section, so that the average phase is either 0 rad. or -1 rad. This is done to allow easy

comparison between different phase distributions. In Figure 5.9 the phases are set to zero wherever $|f_d(x,y)| < 0.06$. The forms of phase$\{\tilde{f}_e^A(x,y)\}$ and phase$\{f_e^B(x,y)\}$ are quite similar to those of phase$\{f_a^A(x,y)\}$ and phase$\{f_a^B(x,y)\}$, respectively, which are described in (5.28). The rms difference between $\Delta\psi_e^{B-A}(x,y)$ and $\Delta\psi^{B-A}(x,y)$, taken over the samples for which $|f_d(x,y)| > 0.06$, is 0.314 rad.

The differences between $\tilde{f}_e^A(x,y)$ and $f_a^A(x,y)$ can be thought of as being noise. Assuming that the statistics of this noise remain the same over the antenna's aperture, it follows that the greater is $|\tilde{f}_e^A(x,y)|$, the closer $\tilde{f}_e^A(x,y)$ approximates $f_a^A(x,y)$. The ratio of $|\tilde{f}_e^A(x,y)|$ at the centre of any speaker to any sample (not located at the centre of a speaker) of $|\tilde{f}_e^A(x,y)|$ is at least 1.5. It is therefore to be expected that the most accurate samples of phase$\{\tilde{f}_e^A(x,y)\}$ are those located at the centres of the speakers. The same reasoning applies to $f_e^B(x,y)$.

Figure 5.10 shows the samples of phase$\{\tilde{f}_e^A(x,y)\}$, phase$\{f_e^B(x,y)\}$ and $\Delta\psi_e^{B-A}(x,y)$ which lie at the centres of the speakers. The rms difference between the samples of $\Delta\psi_e^{B-A}(x,y)$ depicted in Figure 5.10(c) and the corresponding samples of $\Delta\psi^{B-A}(x,y)$ is 0.16 rad.

It is apparent in Figure 5.10(a) that $-$ phase$\{\tilde{f}_e^A(x,y)\}$ tends to increase with increasing x and with decreasing y. Apart from the sample corresponding to speaker 1, the samples of $-$ phase$\{f_e^B(x,y)\}$, depicted in Figure 5.10(b), show the same trend. Because of the translation property of the Fourier transform operator (Table 3.3), this suggests that $A_m^A(u,v)$ is not properly centred. This is confirmed by Figure 5.6(b), which depicts a diagonal cut through $A_m^A(u,v)$. Because $|f_a^A(x,y)|$ is expected to be point symmetric, according to the definition (3.65), $A_m^A(u,v)$ might also be expected to be point symmetric about the point $(0,0)$. However, it is clear from Figure 5.6(b) that $A_m^A(u,v)$ is instead symmetric about a point somewhat translated from $(0,0)$. This suggests that the measurement microphone is mispositioned, in the sense that the direction of a straight line from the centre of the antenna to the measurement microphone is not exactly perpendicular to the aperture plane, when the recorded values of the elevation and azimuth angles are each 0°.

The linear phase term, which best fits the non-zero data presented in Figure 5.10(a) is

$$\psi_{lin}(x,y) = -0.076x + 0.094y \tag{5.31}$$

This implies that $A_m^A(u,v)$ is centred on the point $(0.012/\lambda, -0.015/\lambda)$ rather than, as it should be, on $(0,0)$. This in turn implies that the normal to the aperture plane coincides with the direction of the straight line, from the centre of the antenna to the microphone, when $(\theta_{az};\theta_{el}) = (0.89°; -0.86°)$. To simulate what would have been measured if the microphone had been correctly positioned, $\psi_{lin}(x,y)$ is subtracted from phase$\{\tilde{f}_e^A(x,y)\}$ and phase$\{f_e^B(x,y)\}$. Figure 5.11 depicts the resulting values of the samples of phase$\{\tilde{f}_e^A(x,y)\}$ and phase$\{f_e^B(x,y)\}$ at the centres of the speakers. The rms value of phase$\{\tilde{f}_e^A(x,y)\}$, taken over the samples points at the centres of the speakers, is 0.14 rad. The rms value of [phase$\{f_e^B(x,y)\} - \Delta\psi^{B-A}(x,y)$], taken over the same sample points, is 0.20 rad.

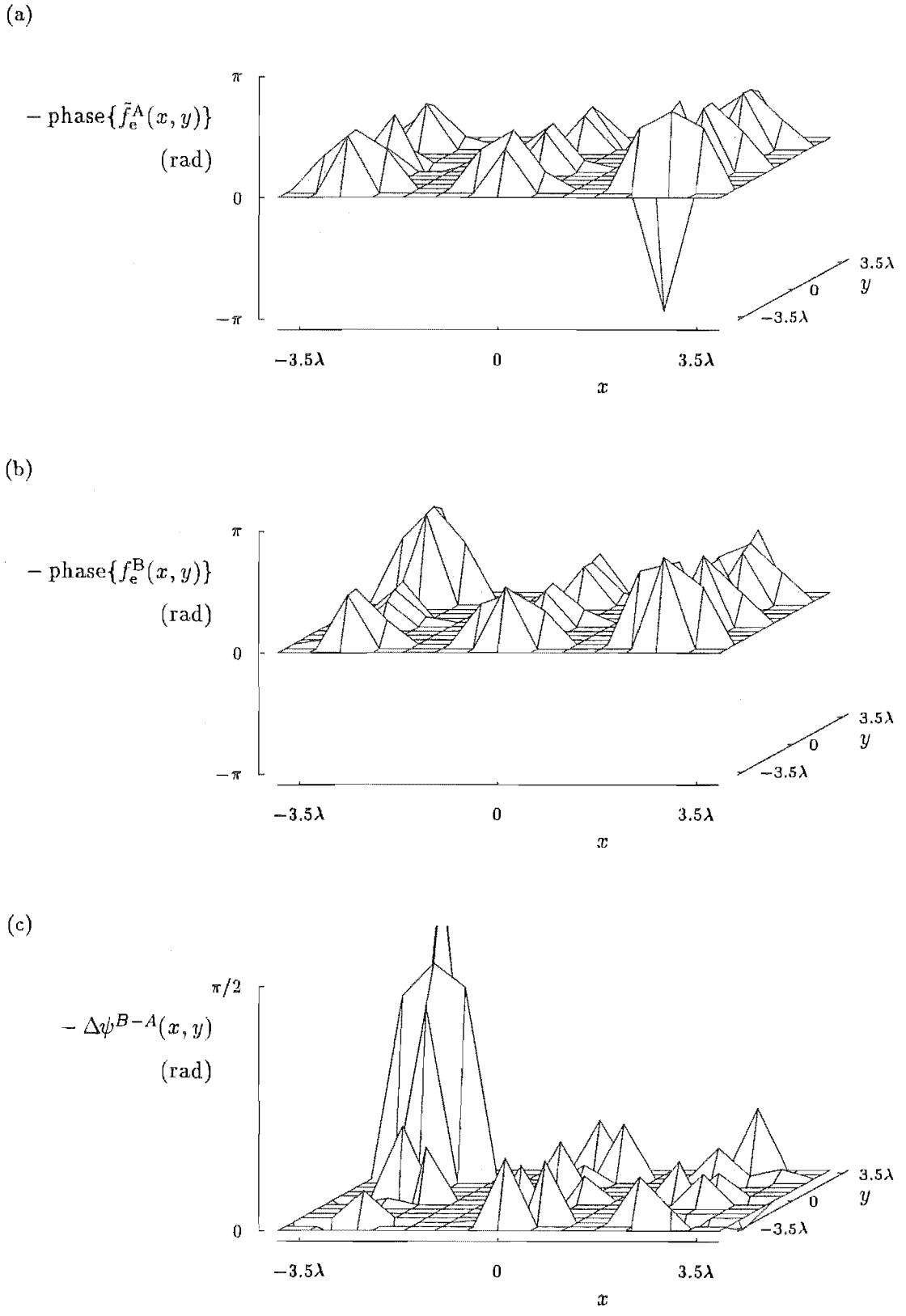
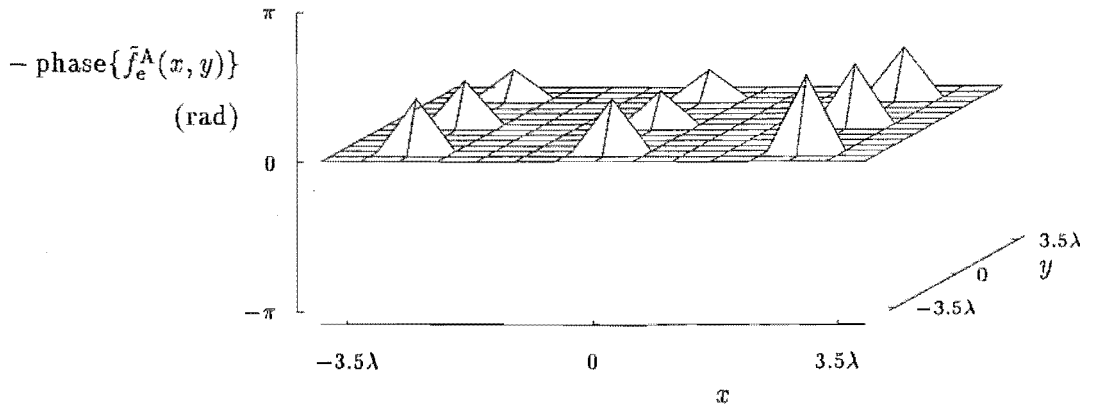
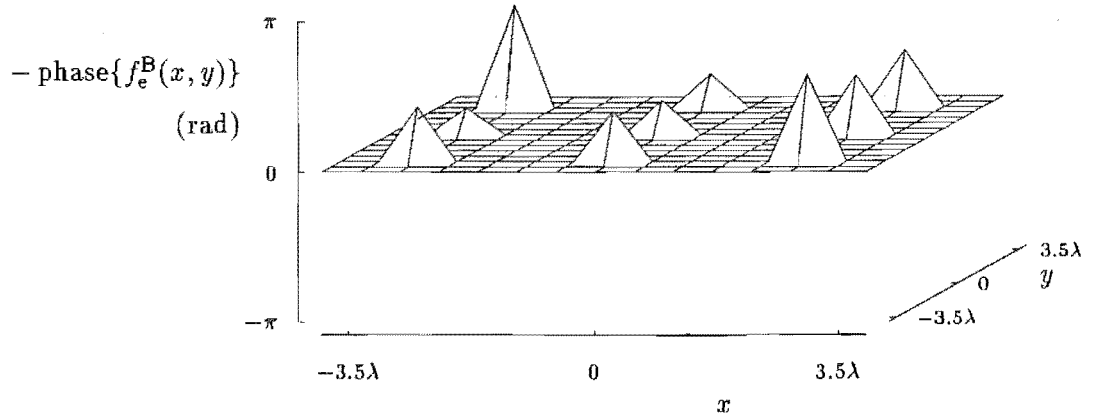


Figure 5.9 Results from modified Gerchberg-Saxton algorithm: (a) phase of $\tilde{f}_e^A(x,y)$; (b) phase of $f_e^B(x,y)$; (c) phase difference between $\tilde{f}_e^A(x,y)$ and $f_e^B(x,y)$. Only the five most central samples are depicted for each speaker.

(a)



(b)



(c)

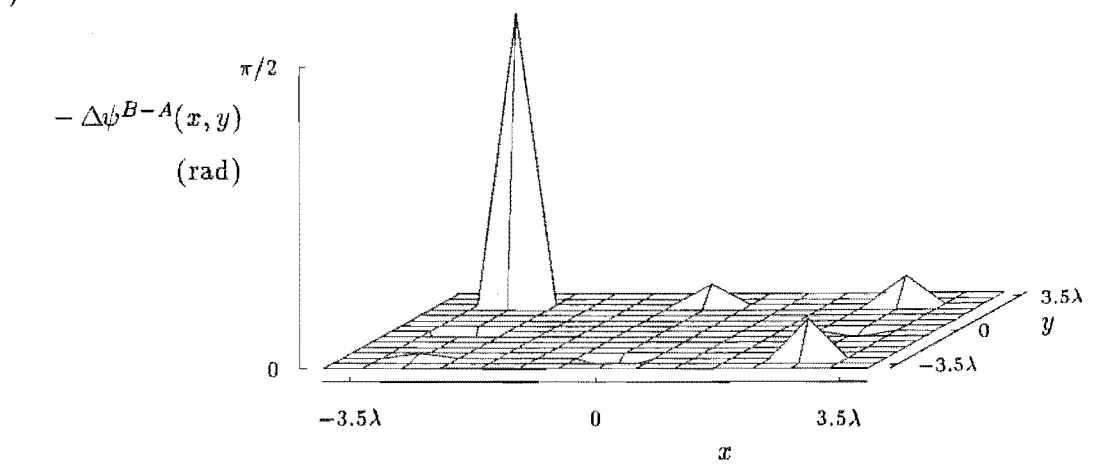
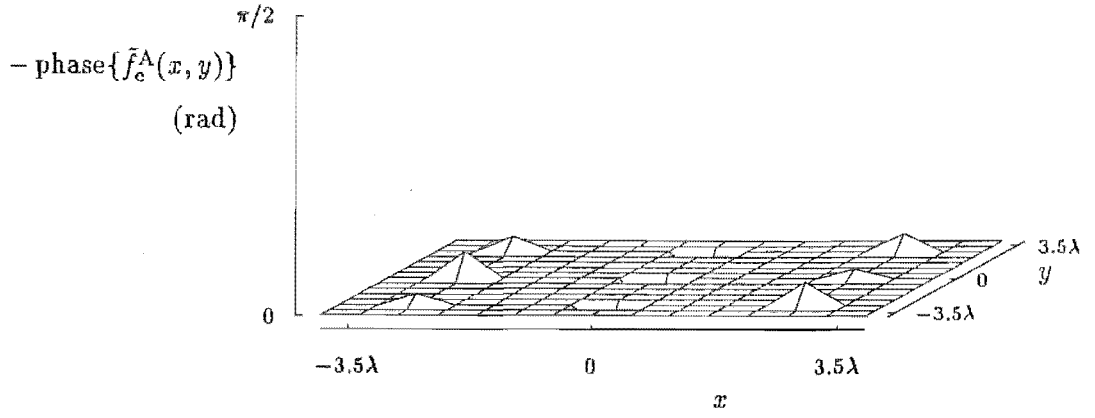


Figure 5.10 Results from modified Gerchberg-Saxton algorithm: (a) phase of $\tilde{f}_e^A(x,y)$; (b) phase of $f_e^B(x,y)$; (c) phase difference between $\tilde{f}_e^A(x,y)$ and $f_e^B(x,y)$. Only the centre sample of each speaker is depicted.

(a)



(b)

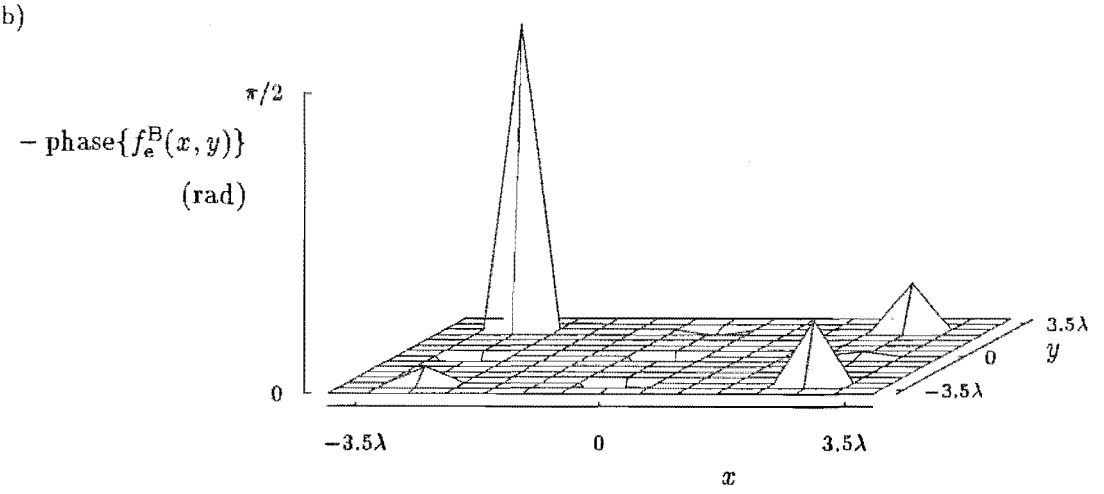


Figure 5.11 Results from modified Gerchberg-Saxton algorithm after a linear phase term has been removed from both phase$\{\tilde{f}_e^A(x,y)\}$ and phase$\{f_e^B(x,y)\}$: (a) phase of $\tilde{f}_e^A(x,y)$; (b) phase of $f_e^B(x,y)$. Only the centre sample of each speaker is depicted.

5.4 SUMMARY

This chapter presents the results of applying the modified Gerchberg-Saxton algorithm to far field amplitude patterns measured in the acoustic laboratory. The measured patterns are those of an acoustic antenna. It is appropriate to measure acoustic fields because, as explained in Section 5.1, they are mathematically analogous to copolar components of electromagnetic fields, under the conditions prevalent in the experiment. In particular, the acoustic aperture field distribution $f_a(x, y)$ and the acoustic weighted far field pattern $F_a(u, v)$ are related by Fourier transformation.

The acoustic antenna is an array of nine speakers, each fed by a signal whose amplitude and phase can be individually adjusted. By setting the amplitudes and phases of the signals fed to the speakers, the antenna can be configured to generate a variety of aperture field distributions. The far field pattern is measured by a microphone. In order to avoid any reflected waves from interfering with the wave propagating directly to the microphone, the signal feeding the antenna consists of a series of tone bursts. The duration of the bursts are short enough so that the directly propagated tone burst reaches the microphone, and its amplitude is measured, before any reflected burst arrives at the microphone. The antenna can be fixed at any elevation angle and is driven, by a motor, about an azimuth axis. A cut through the radiation pattern at any fixed elevation angle is measured by recording, on a pen plotter, the amplitude of the directly propagated tone bursts as a function of azimuth angle. The plotted data of several such constant elevation cuts are input into a computer. Specially written software then converts the elevation and azimuth angles into coordinates on the u, v plane and interpolates the data from the given points in the u, v plane onto a square grid. This enables the data to be operated upon by the modified Gerchberg-Saxton algorithm.

Far field amplitude patterns of two different configurations of the acoustic antenna have been measured. The Fourier transform relationship between $f_a(x, y)$ and $F_a(u, v)$ is predicated on there being no acoustic propagation in directions nearly parallel to the aperture plane. Despite this, however, non-zero fields were measured in such directions. Because they could not have been radiated by the part of the aperture distribution which gave rise to the well collimated part of the far field pattern, the measured data corresponding to far field radiation in directions almost parallel to the aperture plane were set to zero. The measured weighted far field amplitude pattern $A_m(u, v)$ was transformed to provide an estimate of the autocorrelation $ff_m(x, y)$ which, while it is palpably approximate, allows the extent of $f_a(x, y)$ to be usefully estimated. A plausible guess at the form of $|f_a(x, y)|$ can be based on the extent of $f_a(x, y)$ and the physical geometry of the antenna. This guess is taken to be the 'designed' aperture field amplitude distribution $|f_d(x, y)|$ which is utilized by the modified Gerchberg-Saxton algorithm.

The results of applying the modified Gerchberg-Saxton algorithm to the measured data and $|f_d(x, y)|$, for each antenna configuration, are depicted in Figures 5.9 to 5.11. Knowledge of the signals fed to the individual speakers can be invoked to estimate the accuracy to which phase$\{f_e(x, y)\}$, generated by the modified Gerchberg-Saxton algorithm, approximates phase$\{f_a(x, y)\}$. The results indicate that the rms accuracy of phase$\{f_e(x, y)\}$, taken over the sample points at which $|f_d(x, y)| > 0.06$, is 0.3 rad. The results also predicted, to an rms accuracy of within 0.2 rad., the phase of the signals fed to each speaker. These phases were deduced from the samples of phase$\{f_e(x, y)\}$ at the centres of the speakers, after subtracting the linear phase term due (probably)

to the measurement microphone being mispositioned.

The modified Gerchberg-Saxton algorithm was developed to solve the radio engineering phase problem for high gain antennas. However, the results and discussion presented in this chapter indicate that the modified Gerchberg-Saxton algorithm is potentially useful for other antenna types, such as sonic antennas and antenna arrays. In addition, provided the measured far field amplitude pattern is multiplied by an appropriate weighting factor (cf. (5.14)), the algorithm can be applied to antennas which have only moderately high gains.

CHAPTER 6

CONCLUSIONS AND SUGGESTIONS FOR FUTURE WORK

Having completed the research presented in this thesis, it is now obvious to me that there are many ways in which it could be improved and expanded upon. Various suggestions for future work are outlined in Section 6.1. Section 6.2 concludes this thesis with a summary of the main features of the modified Gerchberg-Saxton algorithm.

6.1 SUGGESTIONS FOR FUTURE WORK

In the following sections, several avenues for continuing research are discussed. In Section 6.1.1, suggestions are made for improving the modified Gerchberg-Saxton algorithm. Improved ways of verifying the modified Gerchberg-Saxton algorithm are proposed in Section 6.1.2. The concern in Section 6.1.3 is with how the modified Gerchberg-Saxton algorithm might be invoked for one-dimensional phase retrieval, which could be particularly useful in certain radio engineering applications.

6.1.1 Improvements to the modified Gerchberg-Saxton algorithm

In most of the examples presented in this chapter, the composite algorithm performs as well as can be expected, given the quality of the data to which it is applied. For some of the computer simulations, however, the composite algorithm performs worse than might be expected. An informative example of this is shown in Figure 4.32(a). Ignoring the run for which $\psi_{\mathrm{quad}} = 0.4$ rad., a clear trend in the values of E_{c} and $\mathcal{E}^{\mathrm{ap}}$ is apparent. However, for the run of the composite algorithm corresponding to $\psi_{\mathrm{quad}} = 0.4$ rad., the values of E_{c} and $\mathcal{E}^{\mathrm{ap}}$ far exceed this trend. This suggests that more work might be done to develop a form of the modified Gerchberg-Saxton algorithm which is able to generate the best possible estimate $f_{\mathrm{e}}(x, y)$, of the image-form of $f_{\mathrm{a}}(x, y)$, given any particular set of data $A_{\mathrm{m}}(u, v)$ and $|f_{\mathrm{d}}(x, y)|$.

It is demonstrated in Section 4.4.4 that the convergence properties of the modified Gerchberg-Saxton algorithm improve when the starting aperture field distribution $g_1(x, y)$ is an approximation to $f_{\mathrm{a}}(x, y)$. One way of obtaining such an approximation would be to invoke Davis' method (Sec. 3.5.1) for solving the radio engineering phase problem. This method generates an estimate of the quadratic terms of phase$\{f_{\mathrm{a}}(x, y)\}$ to which phase$\{g_1(x, y)\}$ can then be set. Alternatively, an estimate of $f_{\mathrm{a}}(x, y)$ can be obtained over a grid of, say, 8 by 8 sample points in the aperture plane, by applying the modified Gerchberg-Saxton algorithm to only the centre 8 by 8 samples of $A_{\mathrm{m}}(u, v)$. Note that each iteration of the modified Gerchberg-Saxton algorithm would be completed in a relatively short time, because only a small number of samples are involved. Once an estimate of $f_{\mathrm{a}}(x, y)$, sampled over the 8 by 8 sampling grid, has been generated, it could be interpolated onto a finer grid. The interpolated estimate of $f_{\mathrm{a}}(x, y)$

might then be utilized as $g_1(x,y)$ for the main run of the modified Gerchberg-Saxton algorithm. This technique of starting with a small number of samples and progressing to the full number of samples has been applied in an image processing context by McCallum and Bates [1989] (Sec. 3.4.3.3).

A possible way of increasing the accuracy to which $f_e(x,y)$ approximates the image-form of $f_a(x,y)$ might be to invoke the following averaging technique. Suppose the modified Gerchberg-Saxton algorithm is applied several times to the same $A_m(u,v)$ and $|f_d(x,y)|$. The different runs of the algorithm could utilize different distributions for $g_1(x,y)$, could involve different forms of the modified Gerchberg-Saxton algorithm or could consist of different numbers of iterations. Provided the different estimates $f_e(x,y)$ generated by the different runs of the algorithm are not the same, their average (or the average of their phases) may well be a better estimate of the image-form (or the phase of the image-form) of $f_a(x,y)$ than any of the individual estimates $f_e(x,y)$. Such an averaging technique has been successfully incorporated into the Misell algorithm [Morris, 1985] and with phase retrieval from noisy data [McCallum and Bates, 1989] (Sec. 3.4.3.3).

Two issues which are related to each other are the oversampling of $A_m(u,v)$ and the use of $|f_d(x,y)|$ in the aperture constraint of the modified Gerchberg-Saxton algorithm. These issues are discussed in the following three paragraphs.

Section 4.7 outlines why it is desirable for $A_m(u,v)$ to be oversampled by the smallest factor that still permits the modified Gerchberg-Saxton algorithm to estimate $f_a(x,y)$ usefully accurately. In most of the examples presented in Chapter 4, $A_m(u,v)$ is oversampled by a factor of at least two. However, a preliminary study, presented in Section 4.7.2, suggests that the performance of the composite algorithm may be only slightly degraded when $A_m(u,v)$ is oversampled by a factor of only 1.7.

The accuracy to which $f_e(x,y)$, generated by the modified Gerchberg-Saxton algorithm, approximates the image-form of $f_a(x,y)$ is limited by the accuracy of the data to which the algorithm is applied. Because it is likely that $A_m(u,v)$ approximates $|F_a(u,v)|$ better than $|f_d(x,y)|$ approximates $|f_a(x,y)|$, it is desirable that the accuracy of $f_e(x,y)$ be limited only by the errors in $A_m(u,v)$, so that $|f_d(x,y)|$ is only used where necessary as an aid to the convergence of the algorithm. Unfortunately, this is not true for the composite algorithm. For example, Figure 4.39 shows that $\mathcal{E}^{ap}$ increases with increasing $(|f_a(x,y)| - |f_d(x,y)|)$. The implication is that if $|f_d(x,y)|$ is a relatively bad approximation to $|f_a(x,y)|$, it is a hindrance, rather than a help, to the convergence of the modified Gerchberg-Saxton algorithm. Recall from Sections 4.4.2 and 4.4.3 that the CC and HIO algorithms, which comprise the constant correction algorithm, discard $|f_d(x,y)|$ after 400 out of a total of 500 iterations. The reasons for doing so are discussed in Section 4.4.2. Perhaps a better approach would be to discard $|f_d(x,y)|$ after a relatively small number of iterations if it is a bad approximation to $|f_a(x,y)|$. It must not be forgotten, however, that if $|f_d(x,y)|$ is a good approximation to $|f_a(x,y)|$, it is a help to the convergence of the algorithm and may need not to be discarded at all. So the question is: how does one decide whether $|f_d(x,y)|$ is a sufficiently good or bad approximation?

The discussion in Section 4.7.2 intimates an implicit relationship between the oversampling of $A_m(u,v)$ and the use of $|f_d(x,y)|$ in the aperture constraint. If $|f_d(x,y)|$ is discarded altogether, the modified Gerchberg-Saxton algorithm is effectively attempting to solve the Fourier phase problem defined in (3.41). It is therefore essential that $A_m(u,v)$ be oversampled by a factor of at least two (Sec. 3.4.2.1), because that is a requirement for a solution to the Fourier phase problem to be unique (Sec. 3.4.2.4). On

the other hand, if the modified Gerchberg-Saxton algorithm incorporates a constraint in which $|f_e(x,y)| = |f_d(x,y)|$ and if $|f_d(x,y)|$ is exact, it is highly likely that the algorithm may still converge to a unique solution, even when $|A_m(u,v)|$ is oversampled by a factor of between one and two. Presumably, the more inaccurate is $|f_d(x,y)|$, or the sooner it is discarded by the modified Gerchberg-Saxton algorithm, the greater is the sampling factor required for the algorithm to converge to a unique solution. For reasons intimated in the previous two paragraphs it would be useful to gain a deeper understanding of the interrelationship between the above factors.

6.1.2 Verification of the algorithm

In Chapter 4 the composite algorithm is evaluated by applying it to data generated from a computer model. Such computer simulations are an important part of developing and evaluating the modified Gerchberg-Saxton algorithm because, if the algorithm does not perform well in the idealized environment of a computer model, it is not likely to do so in the real-world either. The computer model described in Section 4.2 could be extended to simulate many further phenomena, such as astigmatic shape defects of reflector surfaces and multiplicative measurement noise. The aim of such extensions would be to make the model more realistic and to test the modified Gerchberg-Saxton algorithm over an even wider range of conditions than are simulated in the examples included in Chapter 4.

In Chapter 5 the modified Gerchberg-Saxton algorithm is applied to data obtained by measuring the amplitude pattern of a sonic antenna. This antenna comprises an array of several speakers, each fed by its own signal. The estimated $f_e(x,y)$, generated by the modified Gerchberg-Saxton algorithm, is accurate enough to confirm that the different speaker aperture field distributions have approximately the same form as each other. The actual phase of the signal feeding each speaker can also be identified from the information contained in $f_e(x,y)$. The results of the experiment would be more useful if the measurement could be made more accurately and if a higher operating frequency could be used. Making more accurate measurements should allow the modified Gerchberg-Saxton algorithm to generate a correspondingly more accurate $f_e(x,y)$ (cf. Sec. 4.8.2). A higher operating frequency would allow the sample points in the aperture plane to be closer together, thereby enabling $f_e(x,y)$ to reveal finer detail in the aperture field distribution.

The main unfinished business is to test the modified Gerchberg-Saxton algorithm on a high gain radio antenna. Demonstrations of the usefulness of the algorithm applied to real-world measurements are necessary if radio engineers are to be persuaded to make use of the algorithm. The point is that a computer model (or a sonic antenna) cannot exactly simulate realistic electromagnetic conditions.

6.1.3 One-dimensional phase retrieval

When commissioning a high gain radio antenna, it is routine to measure its amplitude pattern along one or two cuts [Blake, 1984, Sec. 8.3]. Such measured data are often sufficient to determine whether or not the radiation pattern meets it specifications (cf. (2.75)). It would therefore be very useful to be able to retrieve as much information about $f_a(x,y)$ as possible from measured samples of a cut through $|F_a(u,v)|$.

Consider a cut along the u axis of the far field plane. The actual copolar far field pattern along this cut is a one-dimensional function, here denoted $Q_a(u) = F_a(u,0)$.

From the definition of the inverse Fourier transform operator (Table 3.3), the one-dimensional inverse Fourier transform of $Q_a(u)$ is $q_a(x)$, where

$$q_a(x) = \int f_a(x, y)\, dy \tag{6.1}$$

$q_a(x, y)$ is therefore the *projection* through $f_a(x, y)$. The main problem with working with one-dimensional cuts is that a solution to the one-dimensional Fourier phase problem is non-unique (Sec. 3.4.2.4). This implies that there are many different functions $q_s(x)$ for which $|\mathrm{FT}\{q_s(x)\}| = |Q_a(u)|$.

In one-dimensional phase retrieval, knowledge of the projection $q_d(x, y)$ of $f_d(x, y)$ cannot be utilized in the same way that $f_d(x, y)$ is utilized in two-dimensional phase retrieval. Even if $|f_d(x, y)| = |f_a(x, y)|$, any differences between phase$\{f_d(x, y)\}$ and phase$\{f_a(x, y)\}$ imply that $|q_a(x)| \neq |q_d(x)|$. However, provided that $|f_d(x, y)| = |f_a(x, y)|$, it follows from (6.1) that $|q_a(x)| \leq |q_d(x)|$ for all x. Therefore, any solution $q_s(x)$ can be discounted as incorrect if $|q_s(x)| > |q_d(x)|$ at any x. In this way, the number of solutions might be significantly reduced. This approach requires that the value of, say, $|F_a(0, 0)|$ must be known relative to, say, $|F_d(0, 0)|$ so that $|q_s(x, y)|$ and $|q_d(x)|$ can be scaled by equal amounts. Note that Fright *et al.* [1989] have successfully performed one-dimensional phase retrieval, in the context of acoustic microscopy, by employing Fienup's error reduction algorithm for complex images (Sec. 3.4.3.3). They successfully encourage the algorithm to converge to the correct solution, as opposed to any of the other possible solutions to the one-dimensional Fourier phase problem, by starting the algorithm with a good estimate of $q_a(x)$.

Another approach to determining $f_a(x, y)$ from one or more cuts through $A_m(u, v)$ is similar to the methods described in Section 4.7.3 for dealing with truncated far field data. In both of these situations, $A_m(u, v)$ is not available at all of the points on the far field sampling grid. A possible algorithm would apply a constraint involving $A_m(u, v)$, in the regions of the u, v plane where $A_m(u, v)$ is known. In the remaining regions, $|F_d(u, v)|$ would be utilized to aid with the convergence of the algorithm and to help ensure that any solution is unique. An alternative to utilizing $|F_d(u, v)|$ would be to utilize an interpolation of the known parts of $A_m(u, v)$ onto the region of the u, v plane where $A_m(u, v)$ is not available. Such algorithms would obviously be more successful given more measured cuts through $A_m(u, v)$. Their performance would also be enhanced the more closely $|f_d(x, y)|$ approximates $|f_a(x, y)|$.

6.2 CONCLUSIONS

High gain reflector antennas are suited to applications requiring an antenna whose far field pattern has a high peak gain, a narrow main beam and low sidelobe levels. Any geometrical defects of the antenna tend to degrade the far field pattern (Secs. 3.2 and 2.2.5). It is advantageous to be able to deduce these defects from a single measurement of the copolar far field amplitude pattern. The modified Gerchberg-Saxton algorithm has been developed to achieve this aim.

The purpose of the algorithm is to generate an estimate $f_e(x, y)$ of the actual copolar aperture field distribution $f_a(x, y)$. The geometrical defects of the antenna can then be deduced from phase$\{f_e(x, y)\}$ (Sec. 3.2). Along with the measured copolar far field amplitude pattern $A_m(u, v)$, the modified Gerchberg-Saxton algorithm requires the design copolar aperture amplitude distributions $|f_d(x, y)|$ as an input. The algorithm iterates between the aperture plane and the far field plane, applying constraints in each

plane. The many forms of the modified Gerchberg-Saxton algorithm are characterized by the ways in which the constraints are applied in the two planes.

The particular form of the modified Gerchberg-Saxton algorithm upon which this thesis stands is the composite algorithm (Sec. 4.4.5), which has been evaluated by applying it to computer simulated data (Chap. 4). The results of this evaluation (summarized in Sec. 4.8.4) are encouraging. The performance of the algorithm tends to be independent of the geometrical defects of the antenna. However, the accuracy of phase$\{f_e(x,y)\}$ worsens as either $|f_d(x,y)|$ or $A_m(u,v)$ worsen. Provided that the rms difference between $|F_a(u,v)|$ and $A_m(u,v)$ is less than 50 dB below the peak value of $A_m(u,v)$, and provided that the rms difference between $|f_d(x,y)|$ and $|f_a(x,y)|$ is less than 0.05 times the peak value of $|f_a(x,y)|$, the constant correction algorithm usually generates an $f_e(x,y)$ whose phase approximates that of $f_a(x,y)$ to within an rms value of 0.06 rad. (Sec. 4.8.4).

Chapter 5 provides a practical demonstration of how the modified Gerchberg-Saxton algorithm can be applied in a real-world situation. The composite algorithm is applied to a measured far field amplitude pattern of an acoustic antenna consisting of nine speakers. The phase estimated by the algorithm, phase$\{f_e(x,y)\}$, enabled the phase of the signals feeding the individual speakers to be retrieved to an rms accuracy of within 0.2 rad. (Sec. 5.3.5). This accuracy was achieved in spite of the relatively high estimated noise level of 32 dB below the peak value of the measured far field amplitude pattern (Sec. 5.3.1).

The evaluation of the modified Gerchberg-Saxton algorithm suggests that it may turn out to be a practicable means of inferring geometrical defects of a high gain reflector antenna from only the amplitude of its far field pattern. Other methods for inferring geometrical defects are outlined in Sections 3.3 and 3.5. As a way of comparing these methods with the modified Gerchberg-Saxton approach, the main features of the composite algorithm are listed here:

1. The antenna geometry need not be measured directly (refer to discussion of measurement of reflector shapes in Sec. 3.3.1).

2. The copolar amplitude pattern can be measured in either the Fresnel region or the far field region (cf. near field scanning techniques, Sec. 3.3.2). The measurements can often be conveniently made using a nominally stationary source and by rotating the antenna about two axes. This measurement process is no more complicated than that employed for either complex holography (Sec. 3.3.3.2), the Misell algorithm (Sec. 3.5.3) or the plane-to-plane diffraction algorithm (Sec. 3.5.4). Detailed analysis and discussion in this thesis are predicated on the measurements being made in the far field.

3. No phase measurements are required (unlike complex holography, Sec. 3.3.3.2, and near field scanning techniques, Sec. 3.3.2). Therefore no separate reference antenna is required. Note that phase measurements are suspect at very high frequencies and whenever the measurement source moves unpredictably (cf. (3.11)).

4. Only a single two-dimensional copolar amplitude pattern is required to be measured (the Misell algorithm, Sec. 3.5.3, and the plane-to-plane diffraction algorithm, Sec. 3.5.4, both require two amplitude patterns to be measured). However, in order to resolve the ambiguity mentioned in (8) below, it may be necessary to make extra measurements of the copolar amplitude pattern at a small number of angles.

5. The measurement of $|F_a(u,v)|$ does not require any changes to the geometry of the antenna (unlike the Misell algorithm, Sec. 3.5.3). However, certain changes may be required to resolve the ambiguity mentioned in (8) below.

6. $A_m(u,v)$ must be oversampled by a factor of close to, or greater than, two, over a rectangular grid in the u,v plane. It therefore requires more measured samples in total than does the Misell algorithm (Sec. 3.5.3), the plane-to-plane diffraction algorithm (Sec. 3.5.4), or complex holography (Sec. 3.3.3.2).

7. Design data, or knowledge about $|f_a(x,y)|$, is required to calculate $|f_d(x,y)|$ (Sec. 4.1.1).

8. The estimate $f_e(x,y)$, generated by the modified Gerchberg-Saxton algorithm, suffers from a twofold ambiguity: $f_e(x,y)$ approximates either $f_a(x,y)$ or it approximates $\bar{f}_a(x,y)$. Section 4.1.2 suggests practical means for resolving this ambiguity.

Whether or not these features imply that the modified Gerchberg-Saxton algorithm is preferable to other methods, for determining geometrical defects of a reflector antenna, depends upon the particular application. Factors to be considered include the availability of equipment such as a source antenna, reference antenna and phase measuring equipment.

The main point is that the modified Gerchberg-Saxton algorithm approach is significantly different to other methods and should be taken into consideration when one needs to identify geometrical defects. In situations where it is difficult or inconvenient to measure phase, the modified Gerchberg-Saxton algorithm could well be the method of optimizing the performance of high gain reflector antennas at the least cost.

REFERENCES

ANDERSON, A.P. (1977), 'Microwave holography', *Proceedings of the IEE, IEE Reviews*, Vol. 124, No. 11R, November, pp. 946–962.

ANDERSON, L.J. and GROTH, L.H. (1963), 'Reflector surface deviations in large parabolic antennas', *IEEE Transactions on Antennas and Propagation*, Vol. AP-11, No. 2, March, pp. 148–152.

ANDERSON, A.P. and SALI, S. (1985), 'New possibilities for phaseless microwave diagnostics. Part 1: Error reduction techniques', *IEE Proceedings Part H*, Vol. 132, No. 5, August, pp. 291–298.

ANDERSON, A.P., BENNETT, J.C., WHITAKER, A.J.T. and GODWIN, M.P. (1978), 'Measurement and optimisation of a large reflector antenna by microwave holography', in *International Conference on Antennas and Propagation, Part 1: Antennas*, IEE Conference Publication Number 169, London, November, pp. 128–131.

ANDERSON, A.P., McCORMACK, J.E.M. and JUNKIN, G. (1988), 'Phase retrieval enhancement of antenna metrology data', *Electronics Letters*, Vol. 24, No. 19, September, pp. 1243–1244.

ANDERSON, A.P., JUNKIN, G. and McCORMACK, J.E. (1989), 'Near-field far-field predictions from single-intensity-planar-scan phase retrieval', *Electronics Letters*, Vol. 25, No. 8, April, pp. 519–520.

ARNOLD, J.M. (1986), 'Geometrical theories of wave propagation: a contemporary review', *IEE Proceedings Part J*, Vol. 133, No. 2, April, pp. 165–188.

BAARS, J.W.M. (1973), 'The measurement of large antennas with cosmic radio sources', *IEEE Transactions on Antennas and Propagation*, Vol. AP-21, No. 4, July, pp. 461–474.

BACH, H. and VISKUM, H.-H. (1987), 'The SNFGTD method and its accuracy', *IEEE Transactions on Antennas and Propagation*, Vol. AP-35, No. 2, February, pp. 169–175.

BATES, R.H.T. (1982), 'Astronomical speckle imaging', *Physics Reports (Review Section of Physics Letters)*, Vol. 90, No. 4, October, pp. 203–297.

BATES, R.H.T. (1987a), 'Some image processing: Highlights in retrospect', *Search*, Vol. 18, No. 5, September, pp. 237–240.

BATES, R.H.T. (1987b), 'Twenty years of image processing at Canterbury', *IPENZ Transactions*, Vol. 14, No. 1/EMCh, March, pp. 9–13.

BATES, R.H.T. and McDONNELL, M.J. (1989), *Image restoration and reconstruction*, The Oxford Engineering Science Series; 16, Clarendon Press, Oxford. First published 1986. Corrected and updated 1989.

BATES, R.H.T. and MNYAMA, D. (1986), 'The status of practical Fourier phase retrieval', in HAWKES, P.W. (Ed.), *Advances in Electronics and Electron Physics, Volume 67*, Academic Press Inc., Orlando, pp. 1–64.

BATES, R.H.T. and NAPIER, P.J. (1971), 'A suggestion for determining antenna pattern phase from holographic type of measurement', *Proceedings of the IREE Australia*, Vol. 32, No. 4, April, pp. 164–166.

BATES, R.H.T., FRIGHT, W.R. and GARDENIER, P.H. (1987), 'Gerchberg-Saxton phase retrieval when image magnitude given only approximately', in IDELL, P.S. (Ed.), *Digital Image Recovery and Synthesis*, Proceedings of the SPIE Volume 828, August, pp. 171–176.

BENNETT, J.C. and GODWIN, M.P. (1977), 'Necessary criteria for the diagnosis of panel misalignments in large reflector antennas by microwave metrology', *Electronics Letters*, Vol. 13, No. 16, August, pp. 463–465.

BENNETT, J.C., ANDERSON, A.P., McINNES, P.A. and WHITAKER, A.J.T. (1976), 'Microwave holographic metrology of large reflector antennas', *IEEE Transactions on Antennas and Propagation*, Vol. AP-24, No. 3, May, pp. 295–303.

BERGLAND, G.D. (1969), 'A guided tour of the fast Fourier transform', *IEEE Spectrum*, Vol. 6, No. 7, July, pp. 41–52.

BLAKE, L.V. (1984), *Antennas*, Artech House Inc., Norwood, MA, USA, 2nd ed. Originally published by John Wiley & Sons, 1966.

BORN, M. and WOLF, E. (1970), *Principles of optics*, Pergamon Press, Oxford, 4th ed.

BRACEWELL, R.N. (1978), *The Fourier transform and its applications*, McGraw-Hill International Book Company, Tokyo, 2nd ed.

BRIGHAM, E.O. (1974), *The fast Fourier transform*, Prentice-Hall Inc., Englewood Cliffs, New Jersey.

BRUCK, Yu.M. and SODIN, L.G. (1979), 'On the ambiguity of the image reconstruction problem', *Optics Communications*, Vol. 30, No. 3, September, pp. 304–308.

BYRNE, C.L. and FIDDY, M.A. (1987), 'Estimation of continuous object distributions from limited Fourier magnitude measurements', *Journal of the Optical Society of America A: Optics and Image Science*, Vol. 4, No. 1, January, pp. 112–117.

CCIR (1986a), 'Recomendation 580-1: Radiation diagrams for use as design objectives for antennas of earth stations operating with geostationary satellites', in *Recommendations and Reports of the CCIR, 1986, Volume IV, Part 1: Fixed-Satellite Service*, International Telecommunication Union, Geneva, pp. 136–137.

CCIR (1986b), 'Report 390-5: Earth-station antennas for the fixed-satellite service', in *Recommendations and Reports of the CCIR, 1986, Volume IV, Part 1: Fixed-Satellite Service*, International Telecommunication Union, Geneva, pp. 156–178.

CCIR (1986c), 'Report 998: Performance of small earth-station antennas for the fixed-satellite service', in *Recommendations and Reports of the CCIR, 1986, Volume IV, Part 1: Fixed-Satellite Service*, International Telecommunication Union, Geneva, pp. 178–192.

CEDERQUIST, J.N., FIENUP, J.R., MARRON, J.C. and PAXMAN, R.G. (1988), 'Phase retrieval from experimental far-field speckle data', *Optics Letters*, Vol. 13, No. 8, August, pp. 619–621.

CHRISTIANSEN, W.N. and HÖGBOM, J.A. (1985), *Radiotelescopes*, Cambridge Monographs on Physics, Cambridge University Press, Cambridge, 2nd ed.

CHU, T.S. (1971), 'A note on simulating Fraunhofer radiation patterns in the Fresnel region', *IEEE Transactions on Antennas and Propagation*, Vol. AP-19, No. 5, September, pp. 691–692.

CHU, T.-S. and TURRIN, R.H. (1973), 'Depolarization properties of offset reflector antennas', *IEEE Transactions on Antennas and Propagation*, Vol. AP-21, No. 3, May, pp. 339–345.

CLARRICOATS, P.J.B. and POULTON, G.T. (1977), 'High-efficiency microwave reflector antennas — a review', *Proceedings of the IEEE*, Vol. 65, No. 10, October, pp. 1470–1504.

CLAYDON, B. (1970), 'The effects of phase errors caused by axial displacement of the feed-horn or subreflector in a shaped dual reflector system', in *Conference on Earth Station Technology*, IEE Conference Publication Number 72, London, October, pp. 162–168.

COLLIN, R.E. and ZUCKER, F.J. (1969a), *Antenna theory, Part 1*, Inter-University Electronics Series; 7, McGraw-Hill Book Company, New York.

COLLIN, R.E. and ZUCKER, F.J. (1969b), *Antenna theory, Part 2*, Inter-University Electronics Series; 7, McGraw-Hill Book Company, New York.

COOK, G.G., ANDERSON, A.P. and BENNETT, J.C. (1985), 'Microwave metrology of volumetric antenna structure and performance', in *Fourth International Conference on Antennas and Propagation (ICAP 85)*, University of Warwick, Coventry, UK, April, pp. 203–207.

COOK, G.G., ANDERSON, A.P., WHITAKER, A.J.T. and BENNETT, J.C. (1987), 'High resolution 3D imaging of the current distribution of antennas from far field measurements', in *Fifth International Conference on Antennas and Propagation (ICAP 87), Part 1: Antennas*, York, March, pp. 367–370.

COOK, G.G., ANDERSON, A.P., WHITAKER, A.J.T. and BENNETT, J.C. (1989), 'High resolution three-dimensional microwave imaging of antennas', *IEEE Transactions on Antennas and Propagation*, Vol. 37, No. 6, June, pp. 768–779.

CORNWELL, T.J. and NAPIER, P.J. (1988), 'The focal plane coherence function of an imaging antenna and its use in measuring and correcting aberrations', *Radio Science*, Vol. 23, No. 5, September, pp. 739–748.

CUTLER, C.C. (1947), 'Parabolic-antenna design for microwaves', *Proceedings of the IRE*, Vol. 35, No. 11, November, pp. 1284–1294.

DAINTY, J.C. and FIENUP, J.R. (1987), 'Phase retrieval and image reconstruction for astronomy', in STARK, H. (Ed.), *Image Recovery: Theory and application*, Academic Press Inc., New York, Chap. 7, pp. 231–275.

DAVEY, B.L.K., LANE, R.G. and BATES, R.H.T. (1989), 'Blind deconvolution of noisy complex-valued image', *Optics Communications*, Vol. 69, No. 5–6, January, pp. 353–356.

DAVID, P. and VOGE, J. (1969), *Propagation of waves*, Pergamon Press, Oxford, First English ed. This is a translation of *Propagation des Ondes* published in 1966 by Eyrolles Éditeur, Paris.

DAVIS, J.H. (1970), *The evaluation of reflector antennas*, Electrical Engineering Research Laboratory, The University of Texas at Austin, May. NASA Technical Report No. NGL-006-70-1.

DIJK, J. and MAANDERS, E.J. (1968), 'Optimising the blockcing efficiency in shaped Cassegrain systems', *Electronics Letters*, Vol. 4, No. 18, September, pp. 372–373.

ELLDER, J., LUNDAHL, L. and MORRIS, D. (1984), 'Test of phase-retrieval holography on the Onsala 20 m radiotelescope', *Electronics Letters*, Vol. 20, No. 17, August, pp. 709–710.

EVANS, J.V. (1986), 'Twenty years of international satellite communication', *Radio Science*, Vol. 21, No. 4, July-August, pp. 647–664.

FELDKAMP, G.B. and FIENUP, J.R. (1980), 'Noise properties of images reconstructed from Fourier modulus', in RHODES, W.T. (Ed.), *1980 International Optical Computing Conference*, Proceedings of the SPIE Volume 231, April, pp. 84–93.

FIENUP, J.R. (1978), 'Reconstruction of an object from the modulus of its Fourier transform', *Optics Letters*, Vol. 3, No. 1, July, pp. 27–29.

FIENUP, J.R. (1980), 'Iterative method applied to image reconstruction and to computer-generated holograms', *Optical Engineering*, Vol. 19, No. 3, May, pp. 297–305.

FIENUP, J.R. (1982), 'Phase retrieval algorithms: a comparison', *Applied Optics*, Vol. 21, No. 15, August, pp. 2758–2769.

FIENUP, J.R. (1987), 'Reconstruction of a complex-valued object from the modulus of its Fourier transform using a support constraint', *Journal of the Optical Society of America A: Optics and Image Science*, Vol. 4, No. 1, January, pp. 118–123.

FIENUP, J.R. and WACKERMAN, C.C. (1987), 'Phase-retrieval stagnation problems and solutions', *Journal of the Optical Society of America A: Optics and Image Science*, Vol. 3, No. 11, November, pp. 1897–1907.

FINDLAY, J.W. (1964), 'Radio telescopes', *IEEE Transactions on Antennas and Propagation*, Vol. AP-12, No. 7, December, pp. 853–864.

FINDLAY, J.W. (1971), 'Filled-aperture antennas for radio astronomy', *Annual Review of Astronomy and Astrophysics*, Vol. 9, pp. 271–292.

FOURIKIS, N. (1988), 'A parametric study of the constraints related to Gregorian/Cassegrain offset reflectors having negligible cross polarization', *IEEE Transactions on Antennas and Propagation*, Vol. 36, No. 1, January, pp. 144–147.

FRIGHT, W.R., BATES, R.H.T., ROWE, J.M., SPENCER, D.S., SOMEKH, M.G. and BRIGGS, G.A.D. (1989), 'Reconstruction of the complex reflectance function in acoustic microscopy', *Journal of Microscopy*, Vol. 153, Pt 1, January, pp. 103–117.

GABOR, D. (1948), 'A new microscopic principle', *Nature*, Vol. 161, No. 4098, May, pp. 777–778.

GABOR, D. (1949), 'Microscopy by reconstructed wave-fronts', *Proceedings of the Royal Society of London: Series A. Mathematical and Physical Sciences*, Vol. 197, July, pp. 454–487.

GALINDO, V. (1964), 'Design of dual-reflector antennas with arbitrary phase and amplitude distributions', *IEEE Transactions on Antennas and Propagation*, Vol. AP-12, No. 4, July, pp. 403–408.

GALINDO-ISRAEL, V., IMBRIALE, W.A. and MITTRA, R. (1987), 'On the theory of the synthesis of single and dual offset shaped reflector antennas', *IEEE Transactions on Antennas and Propagation*, Vol. AP-35, No. 8, August, pp. 887–896.

GALLOIS, P. (1987), 'Life expectancy of communication satellites', *Electronics & Power*, Vol. 33, No. 9, September, pp. 547–550.

GARDENIER, P.H., McCALLUM, B.C. and BATES, R.H.T. (1986a), 'Fourier transform magnitudes are unique pattern recognition templates', *Biological Cybernetics*, Vol. 54, No. 6, September, pp. 385–391.

GARDENIER, P.H., LIM, C.A., TAN, D.G.H. and BATES, R.H.T. (1986b), 'Aperture distribution phase from single radiation pattern measurement via Gerchberg-Saxton algorithm', *Electronics Letters*, Vol. 22, No. 2, January, pp. 113–115.

GARDENIER, P.H., LIM, C.A., TAN, D.G.H. and BATES, R.H.T. (1986c), 'Feed-position and reflector-shape errors of satellite communications antenna from radiation pattern magnitude', in *IPENZ Conference 86*, Auckland University, New Zealand, February.

GARDENIER, P.H., LIM, C.A. and PARKER, C.R. (1988), 'Satellite communications antenna misalignments inferred from far field magnitude', in *Proceedings of the 25th New Zealand National Electronics Conference*, NELCON, Christchurch, August, pp. 83–88.

GERCHBERG, R.W. (1974), 'Super-resolution through error energy reduction', *Optica Acta*, Vol. 21, No. 9, September, pp. 709–720.

GERCHBERG, R.W. (1986), 'The lock problem in the Gerchberg-Saxton algorithm for phase retrieval', *Optik*, Vol. 74, No. 3, October, pp. 91–93.

GERCHBERG, R.W. and SAXTON, W.O. (1972), 'A practical algorithm for the determination of phase from image and diffraction plane pictures', *Optik*, Vol. 35, No. 2, April, pp. 237–246.

GHOBRIAL, S.I. (1979), 'Off-axis cross-polarization and polarization efficiencies of reflector antennas', *IEEE Transactions on Antennas and Propagation*, Vol. AP-27, No. 4, July, pp. 460–466.

GODWIN, M.P., ANDERSON, A.P. and BENNETT, J.C. (1978), 'Optimisation of feed position and improved profile mapping of a reflector antenna from microwave holographic measurements', *Electronics Letters*, Vol. 14, No. 5, March, pp. 134–136.

GODWIN, M.P., WHITAKER, A.J.T., BENNETT, J.C. and ANDERSON, A.P. (1981), 'Microwave diagnostics of the Chilbolton 25m antenna using the OTS satellite', in *Second International Conference on Antennas and Propagation, Part 1: Antennas*, IEE Conference Publication Number 195, University of York, York, UK, April, pp. 232–236.

GODWIN, M.P., HENCHY, F.B.N. and MARSHALL, W.N. (1985), 'Design and validation of a corrective subreflector for a 13.7 m antenna by microwave holography', *IEE Proceedings Part H*, Vol. 132, No. 7, December, pp. 447–450.

GODWIN, M.P., SCHOESSOW, E.P. and GRAHL, B.H. (1986), 'Improvement of the Effelsberg 100 meter telescope based on holographic reflector surface measurement', *Astronomy and Astrophysics*, Vol. 167, No. 2, pp. 390–394.

GOODMAN, J.W. (1968), *Introduction to Fourier optics*, McGraw-Hill Book Company, San Fransisco.

GREEN, H.E. (1983), 'Antenna pattern measurement with a geostationary satellite', *Journal of Electrical and Electronics Engineering, Australia*, Vol. 3, No. 1, March, pp. 8–17.

HALL, M.P.M. (1979), *Effects of the troposphere on radio communication*, IEE Electromagnetic Waves Series; 8, Peter Peregrinus Ltd., Stevenage, UK.

HANNAN, P.W. (1961), 'Microwave antennas derived from the Cassegrain telescope', *IRE Transactions on Antennas and Propagation*, Vol. AP-9, No. 2, March, pp. 140–153.

HANSEN, P.C. and LARSEN, F.H. (1984), 'Suppression of reflections by directive probes in spherical near-field measurements', *IEEE Transactions on Antennas and Propagation*, Vol. AP-32, No. 2, February, pp. 119–125.

HASSELMANN, F.J.V. and FELSEN, L.B. (1982), 'Asymptotic analysis of parabolic reflector antennas', *IEEE Transactions on Antennas and Propagation*, Vol. AP-30, No. 4, July, pp. 677–685.

HAYES, M.H. (1982), 'The reconstruction of a multidimensional sequence from the phase or magnitude of its Fourier transform', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-30, No. 2, April, pp. 140–154.

HAYES, M.H. and McCLELLAN, J.H. (1982), 'Reducible polynomials in more than one variable', *Proceedings of the IEEE*, Vol. 70, No. 2, February, pp. 197–198.

von HOERNER, S. (1976), 'The design of correcting secondary reflectors', *IEEE Transactions on Antennas and Propagation*, Vol. AP-24, No. 3, May, pp. 336–340.

von HOERNER, S. and WONG, W.-Y. (1975), 'Gravitational deformation and astigmatism of tiltable radio telescopes', *IEEE Transactions on Antennas and Propagation*, Vol. AP-23, No. 5, September, pp. 689–695.

IEEE (1979), *IEEE Standard test procedures for antennas*, IEEE Inc. Distributed in cooperation with Wiley-Interscience.

IEEE (1984), *IEEE Standard dictionary of electrical and electronics terms*, IEEE Inc., New York, 3rd ed. F. Jay (Ed.). Distributed in cooperation with Wiley-Interscience.

INTERSIL (1981), *Data book 1981*, Intersil Inc., California.

ITT (1968), *Reference data for radio engineers*, Howard W. Sams & Co. Inc. (A subsidiary of International Telephone and Telegraph Corporation), Indianapolis, 5th ed.

JAMES, G.L. (1980), 'Analysis of radiation pattern and G/T_A for shaped dual-reflector antennas', *IEE Proceedings Part H*, Vol. 127, No. 1, February, pp. 52–60.

JAMES, G.L. (1986), *Geometrical theory of diffraction for electromagnetic waves*, IEE Electromagnetic waves series; 1, Peter Peregrinus Ltd., London, 3rd ed.

JAMES, G.L. and KERDEMELIDIS, V. (1973), 'Reflector antenna radiation pattern analysis by equivalent edge currents', *IEEE Transactions on Antennas and Propagation*, Vol. AP-21, No. 1, January, pp. 19–24.

JASIK, H. (Ed.) (1961), *Antenna engineering handbook*, McGraw-Hill Book Company Inc., New York, 1st ed.

JOHNSON, R.C., ECKER, H.A. and HOLLIS, J.S. (1973), 'Determination of far-field antenna patterns from near-field measurements', *Proceedings of the IEEE*, Vol. 61, No. 12, December, pp. 1668–1694.

JONES, E.M.T. (1954), 'Paraboloid reflector and hyperboloid lens antennas', *IRE Transactions on Antennas and Propagation*, Vol. AP-2, July, pp. 119–127.

JORDAN, E.C. and BALMAIN, K.G. (1968), *Electromagnetic waves and radiating systems*, Prentice-Hall Electrical Engineering Series, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 2nd ed.

JULL, E.V. (1981), *Aperture antennas and diffraction theory*, IEE Electromagnetic Waves Series; 10, Peter Peregrinus Ltd, Stevenage, UK.

KALCINA, A.P., HAWKE, R.D. and NEIL, G.T. (1987), 'Optimising satellite earth station antennas by microwave holography', *Journal of Electrical and Electronics Engineering, Australia*, Vol. 7, No. 4, December, pp. 267–270.

KELLY, J.B., NEATE, P.R. and SHINN, D.H. (1970), 'Setting and checking the position of the surface of a large reflector', in *Conference on Earth Station Technology*, IEE Conference Publication Number 72, London, October, pp. 227–232.

KERBYSON, P., ANDERSON, A.P. and BENNETT, J.C. (1987), 'Effect of support strut diffraction on reflector surface profile assessment', *Electronics Letters*, Vol. 23, No. 4, February, pp. 146–147.

KINSLER, L.E., FREY, A.R., COPPENS, A.B. and SANDERS, J.V. (1982), *Fundamentals of acoustics*, John Wiley & Sons, New York, 3rd ed.

KO, W.L., MITTRA, R. and LEE, S.W. (1984), 'Aperture blockage in reflector antennas', *IEEE Transactions on Antennas and Propagation*, Vol. AP-32, No. 3, March, pp. 282–287.

KUMMER, W.H. and GILLESPIE, E.S. (1978), 'Antenna measurements—1978', *Proceedings of the IEEE*, Vol. 66, No. 4, April, pp. 483–507.

LAMB, J.W. and OLVER, A.D. (1986), 'Blockage due to subreflector supports in large radiotelescope antennas', *IEE Proceedings Part H*, Vol. 133, No. 1, February, pp. 43–49.

LANE, R.G. (1987), 'Recovery of complex images from Fourier magnitude from phase', *Optics Communications*, Vol. 63, No. 1, July, pp. 6–10.

LANE, R. (1988), *Blind deconvolution and phase retrieval*, Ph.D. thesis, University of Canterbury, Christchurch, New Zealand, February.

LANE, R.G. and BATES, R.H.T. (1987a), 'Automatic multidimensional deconvolution', *Journal of the Optical Society of America A: Optics and Image Science*, Vol. 4, No. 1, January, pp. 180–188.

LANE, R.G. and BATES, R.H.T. (1987b), 'Relevance for blind deconvolution of recovering Fourier magnitude', *Optics Communications*, Vol. 63, No. 1, July, pp. 11–14.

LANE, R.G., FRIGHT, W.R. and BATES, R.H.T. (1987), 'Direct phase retrieval', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-35, No. 4, April, pp. 520–526.

LEE, S.-W. (1977), 'Comparison of uniform asymptotic theory and Ufimtsev's theory of electromagnetic edge diffraction', *IEEE Transactions on Antennas and Propagation*, Vol. AP-25, No. 2, March, pp. 162–170.

LEVY, G.S., BATHKER, D.A., LUDWIG, A.C., NEFF, D.E. and SEIDEL, B.L. (1967), 'Lunar range radiation patterns of a 210-foot antenna at S-band', *IEEE Transactions on Antennas and Propagation*, Vol. AP-15, No. 2, March, pp. 311–313.

LUDWIG, A.C. (1973), 'The definition of cross polarization', *IEEE Transactions on Antennas and Propagation*, Vol. AP-21, No. 1, January, pp. 116–119.

MA, M.T. (1974), *Theory and application of antenna arrays*, John Wiley & Sons, New York. A Wiley-Interscience Publication.

MAKHOUL, J. (1975), 'Linear prediction: a tutorial review', *Proceedings of the IEEE*, Vol. 63, No. 4, April, pp. 561–580.

MATSUNAKA, N., BETSUDAN, S. and KATAGI, T. (1981), 'Sidelobe level reduction by improvement of strut shape', in *AP-S International Symposium Digest, Vol. 2*, IEEE Antennas and Propagation Society, Los Angeles, June, pp. 496–499.

MAXWELL, J.C. (1865), 'A dynamical theory of the electromagnetic field', *Royal Society Transactions*, Vol. CLV. Reprinted in Niven W.D. (Ed.) (1890), *The scientific papers of James Clerk Maxwell, Volume I*, The University Press, Cambridge, pp. 526–597.

MAYER, C.E., DAVIS, J.H., PETERS, III, W.L. and VOGEL, W.J. (1983), 'A holographic surface measurement of the Texas 4.9-m antenna at 86 GHz', *IEEE Transactions on Instrumentation and Measurement*, Vol. IM-32, No. 1, March, pp. 102–109.

McCALLUM, B.C. and BATES, R.H.T. (1989), 'Towards a strategy for automatic phase retrieval from noisy Fourier intensities', *Journal of Modern Optics*, Vol. 36, No. 5, May, pp. 619–648.

McCALLUM, B.C., GARDENIER, P.H. and BATES, R.H.T. (1986), 'Invertible invariant transformations for robotic catalogues', in *Proceedings of the International Conference on Future Computing Systems*, Christchurch, New Zealand, February, pp. 151–158.

McCORMACK, J.E. and ANDERSON, A.P. (1988), 'Phase retrieval microwave metrology of reflector antennas from a single amplitude data set', in *Proceedings of the 11th ESTEC workshop on antenna measurements*, Gothenburg, Sweden, June, pp. 289–294.

McCORMACK, J.E., JUNKIN, G., ANDERSON, A.P. and WHITAKER, A.J.T. (1989), 'Microwave antenna metrology from a single intensity scan', in *Sixth International Conference on Antennas and Propagation (ICAP 89), Part 1: Antennas*, IEE Conference Publication Number 301, Coventry, UK, April, pp. 468–472.

McGRANE, A.R. (1983), 'Range limitations in the computation of antenna far field patterns from Fresnel field measurements', in *Third International Conference on Antennas and Propagation (ICAP 83), Part 1: Antennas*, IEE Conference Publication Number 219, Norwich, April, pp. 515–519.

MILLMAN, J. (1979), *Microelectronics: Digital and analog circuits and systems*, McGraw-Hill International Book Company, Tokyo.

MILNER, M.O. and BATES, R.H.T. (1980), 'Design of subreflectors to compensate for Cassegrain main reflector deformations', *IEE Proceedings Part H*, Vol. 127, No. 5, October, pp. 277–281.

MILNER, M.O., GARDENIER, P.H. and BATES, R.H.T. (1987), 'Antenna aperture phase from far field magnitude', in *IEEE/AP-S International Symposium and URSI Radio Science Meeting*, Virginia Tech, Blacksburg, Virginia, USA, June.

MINARD, R.A., ROBINSON, B.S. and BATES, R.H.T. (1985), 'Full-wave computed tomography. Part 3: Coherent shift-and-add imaging', *IEE Proceedings Part A*, Vol. 132, No. 1, January, pp. 50–58.

MISELL, D.L. (1973a), 'A method for the solution of the phase problem in electron microscopy', *Journal of Physics D: Applied Physics*, Vol. 6, No. 18, December, pp. L6–L9.

MISELL, D.L. (1973b), 'An examination of an iterative method for the solution of the phase problem in optics and electron optics: I. Test calculations', *Journal of Physics D: Applied Physics*, Vol. 6, No. 18, December, pp. 2200–2216.

MISELL, D.L. (1973c), 'An examination of an iterative method for the solution of the phase problem in optics and electron optics: II. Sources of error', *Journal of Physics D: Applied Physics*, Vol. 6, No. 18, December, pp. 2217–2225.

MISELL, D.L. (1978), 'The phase problem in electron microscopy', in COSSLETT, V.E. and BARER, R. (Eds.), *Advances in optical and electron microscopy, Volume 7*, Academic Press, London, pp. 185–279.

MITTRA, R., IMBRIALE, W.A. and MAANDERS, E.J. (Eds.) (1983), *Satellite communication antenna technology*, North-Holland, Amsterdam.

MIYA, K. (Ed.) (1981), *Satellite communications technology*, KDD Engineering and Consulting Inc., Tokyo.

MORRIS, D. (1985), 'Phase retrieval in the radio holography of reflector antennas and radio telescopes', *IEEE Transactions on Antennas and Propagation*, Vol. AP-33, No. 7, July, pp. 749–755. See also correction in Vol. AP-33, p. 1419.

MORRIS, D., HEIN, H., STEPPE, H. and BAARS, J.W.M. (1988), 'Phase retrieval radio holography in the Fresnel region: Tests on the 30 m telescope at 86 GHz', *IEE Proceedings Part H*, Vol. 135, No. 1, February, pp. 61–64.

MOSTOWSKI, A. and STARK, M. (1964), *Introduction to higher algebra*, Pergamon Press, Oxford. Translated from the Polish by MUSIELAK J.

NAPIER, P.J. and BATES, R.H.T. (1971), 'Holographic approach to radiation pattern measurement — II Experimental verification', *International Journal of Engineering Science*, Vol. 9, No. 12, December, pp. 1193–1208.

NAPIER, P.J. and BATES, R.H.T. (1973), 'Antenna — aperture distributions from holographic type of radiation-pattern measurement', *Proceedings of the IEE*, Vol. 120, No. 1, January, pp. 30–34.

NAPIER, P.J., THOMPSON, A.R. and EKERS, R.D. (1983), 'The very large array: design and performance of a modern synthesis radio telescope', *Proceedings of the IEEE*, Vol. 71, No. 11, November, pp. 1295–1320.

NATIONAL (1980), *Linear applications handbook*, National Semiconductor Corporation, California.

NATIONAL (1981), *CMOS databook*, National Semiconductor Corporation, California.

O'HARA, J.G. and PRICHA, W. (1987), *Hertz and the Maxwellians*, IEE History of Technology Series: 8, Peter Peregrinus Ltd, London.

OPPENHEIM, A.V. and SCHAFER, R.W. (1975), *Digital signal processing*, Prentice-Hall Inc., Englewood Cliffs, New Jersey.

OTT, J.H. and RICE, J.S. (1986), 'Sonic microwave-antenna simulator', *IEEE Transactions on Antennas and Propagation*, Vol. AP-34, No. 12, December, pp. 1418–1424.

PANARELLA, E. (1985), 'Diffraction of statistically independent photons from a laser source: photon counting on the diffraction pattern', *Speculations in Science and Technology*, Vol. 8, No. 1, pp. 35–49.

PANARELLA, E. and GUTY, V. (1983), 'Diffraction of statistically independent photons from a laser source: detection of the diffraction pattern with a high-gain photomultiplier', *Speculations in Science and Technology*, Vol. 6, No. 4, pp. 383–390.

PANTER, P.F. (1972), *Communication systems design: Line-of-sight and tropo-scatter systems*, McGraw-Hill Book Company, New York.

PAPI, G., RUSSO, V. and SOTTINI, S. (1971), 'Microwave holographic interferometry', *IEEE Transactions on Antennas and Propagation*, Vol. AP-19, No. 6, November, pp. 740–746.

PAPOULIS, A. (1975), 'A new algorithm in spectral analysis and band-limited extrapolation', *IEEE Transactions on Circuits and Systems*, Vol. CAS-22, No. 9, September, pp. 735–742.

PARINI, C.G., LAU, A.K.K. and CLARRICOATS, P.J.B. (1989), 'Phase-only reflector antenna metrology', *IEE Proceedings Part H*, Vol. 136, No. 4, August, pp. 343–349.

PARIS, D.T., LEACH, JR., W.M. and JOY, E.B. (1978), 'Basic theory of probe-compensated near-field measurements', *IEEE Transactions on Antennas and Propagation*, Vol. AP-26, No. 3, May, pp. 373–379.

PAYNE, J.M. (1973), 'An optical distance measuring instrument', *The Review of Scientific Instruments*, Vol. 44, No. 3, March, pp. 304–306.

PAYNE, J.M., HOLLIS, J.M. and FINDLAY, J.W. (1976), 'New method of measuring the shape of precise antenna reflectors', *The Review of Scientific Instruments*, Vol. 47, No. 1, January, pp. 50–55.

PEARSON, T.J. and READHEAD, A.C.S. (1984), 'Image formation by self-calibration in radio astronomy', *Annual Review of Astronomy and Astrophysics*, Vol. 22, pp. 97–130.

PICQUENARD, A. (1974), *Radio wave propagation*, The Macmillan Press Ltd, London. Copyright N.V. Philips' Gloeilampenfabrieken, Eindhoven, 1974.

RAHMAT-SAMII, Y. (1984), 'Surface diagnosis of large reflector antennas using microwave holographic metrology: An iterative approach', *Radio Science*, Vol. 19, No. 5, September, pp. 1205–1217.

RAHMAT-SAMII, Y. (1985), 'Microwave holography of large reflector antennas — simulation algorithms', *IEEE Transactions on Antennas and Propagation*, Vol. AP-33, No. 11, November, pp. 1194–1203. See also correction in Vol. AP-34, p. 853.

RAHMAT-SAMII, Y. (1987), 'Microwave holographic metrology for antenna diagnosis', *IEEE Antennas and Propagation Society Newsletter*, Vol. 29, No. 3, June, pp. 5–15.

RAHMAT-SAMII, Y. and CHEUNG, R.L.-T. (1987), 'Nonuniform sampling techniques for antenna applications', *IEEE Transactions on Antennas and Propagation*, Vol. AP-35, No. 3, March, pp. 268–279.

RAHMAT-SAMII, Y. and LEMANCZYK, J. (1988), 'Application of spherical near-field measurenemts to microwave holographic diagnosis of antennas', *IEEE Transactions on Antennas and Propagation*, Vol. 36, No. 6, June, pp. 869–878.

RAMO, S., WHINNERY, J.R. and VAN DUZER, T. (1965), *Fields and waves in communication electronics*, John Wiley & Sons, New York.

RANSOM, P.L. and MITTRA, R. (1971), 'A method of locating defective elements in large phased arrays', *Proceedings of the IEEE*, Vol. 59, No. 6, June, pp. 1029–1030.

REITBOECK, H.J. and ALTMANN, J. (1984), 'A model for size- and rotation-invariant pattern processing in the visual system', *Biological Cybernetics*, Vol. 51, No. 2, November, pp. 113–121.

REPJAR, A.G. and KREMER, D.P. (1982), 'Accurate evaluation of a millimeter wave compact range using planar near-field scanning', *IEEE Transactions on Antennas and Propagation*, Vol. AP-30, No. 3, May, pp. 419–425.

RETICON (1983), *Analog signal processing products*, RECTICON corporation, USA.

ROSSI, M. (1988), *Acoustics and electroacoustics*, Artech House, Norwood, USA. This is a translation of *Electroacoustique*, published in 1986 by Presses Polytechniques Romandes, Lausanne, Switzerland.

RUDGE, A.W. and ADATIA, N.A. (1978), 'Offset-parabolic-reflector antennas: a review', *Proceedings of the IEEE*, Vol. 66, No. 12, December, pp. 1592–1618.

RUDGE, A.W. and DAVIES, D.E.N. (1970), 'Electronically controllable primary feed for profile-error compensation of large parabolic reflectors', *Proceedings of the IEE*, Vol. 117, No. 2, February, pp. 351–358.

RUDGE, A.W., MILNE, K., OLVER, A.D. and KNIGHT, P. (Eds.) (1982), *The handbook of antenna design, Volume 1*, IEE Electromagnetic Waves Series; 15, Peter Peregrinus Ltd, London.

RUMSEY, V.H., DESCHAMPS, G.A., KALES, M.L. and BOHNERT, J.I. (1951), 'Techniques for handling elliptically polarized waves with special reference to antennas', *Proceedings of the IRE*, Vol. 39, No. 5, May, pp. 533–556.

RUSCH, W.V.T. (1984), 'The current state of the reflector antenna art', *IEEE Transactions on Antennas and Propagation*, Vol. AP-32, No. 4, April, pp. 313–329.

RUSCH, W.V.T. and POTTER, P.D. (1970), *Analysis of reflector antennas*, Electrical Science Series, Academic Press, New York.

RUSCH, W.V.T., SØRENSEN, O. and BAARS, J.W.M. (1982), 'Radiation cones from feed-support struts of symmetric paraboloidal antennas', *IEEE Transactions on Antennas and Propagation*, Vol. AP-30, No. 4, July, pp. 786–789.

RUZE, J. (1965), 'Lateral-feed displacement in a paraboloid', *IEEE Transactions on Antennas and Propagation*, Vol. AP-13, No. 5, September, pp. 660–665.

RUZE, J. (1966), 'Antenna tolerance theory — a review', *Proceedings of the IEEE*, Vol. 54, No. 4, April, pp. 633–640.

SAFAK, M. and DELOGNE, P.P. (1976), 'Cross polarization in Cassegrainian and front-fed paraboloidal antennas', *IEEE Transactions on Antennas and Propagation*, Vol. AP-24, No. 4, July, pp. 497–501.

SALI, S. (1988), 'Phase-retrieval technique for antenna metrology', *Electronics Letters*, Vol. 24, No. 2, January, pp. 132–133.

SALI, S. and ANDERSON, A.P. (1987a), 'Experimental investigation of 'plane-to-plane' diffraction phase retrieval for antenna metrology', in *Fifth International Conference on Antennas and Propagation (ICAP 87), Part 1: Antennas*, IEE Conference Publication Number 274, York, March, pp. 399–401.

SALI, S. and ANDERSON, A.P. (1987b), 'Uncertainty in deriving phase information from far-field modulus only', *Electronics Letters*, Vol. 23, No. 8, April, pp. 426–428.

SANZ, J.L.C. and HUANG, T.S. (1985), 'Polynomial system of equations and its applications to the study of the effect of noise on multidimensional Fourier transform phase retrieval from magnitude', *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-33, No. 4, August, pp. 997–1004.

SAXTON, W.O. (1978), *Computer techniques for image processing in electron microscopy*, Advances in Electronics and Electron Physics; 10, Academic Press, New York.

SCOTT, P.F. and RYLE, M. (1977), 'A rapid method for measuring the figure of a radio telescope reflector', *Monthly Notices of the Royal Astronomical Society*, Vol. 178, No. 3, March, pp. 539–545.

SHANKLIN, J.P. (1955), 'Pattern measurements of large fixed antennas', *IRE Transactions on Instrumentation*, Vol. PGI-4, October, pp. 16–22.

SHEWELL, J.R. and WOLF, E. (1968), 'Inverse diffraction and a new reciprocity theorem', *Journal of the Optical Society of America*, Vol. 58, No. 12, December, pp. 1596–1603.

SILVER, S. (Ed.) (1949), *Microwave antenna theory and design*, MIT Radiation Laboratory Series; 12, McGraw-Hill Book Company Inc. Republished in 1965 by Dover Publications Inc., New York.

SINTON, A.M., GARDENIER, P.H. and BATES, R.H.T. (1986), 'Reinvestigation of optical interference at low light levels', *Speculations in Science and Technology*, Vol. 9, No. 4, November, pp. 269–278.

SLATER, R.H. (1970), 'Radiation pattern of imperfect paraboloidal reflectors', *Electronics Letters*, Vol. 6, No. 25, December, pp. 796–798.

SLATER, R.H. (1971), 'Metrology and radio performance of reflector of Chilbolton aerial', *Proceedings of the IEE*, Vol. 118, No. 12, December, pp. 1691–1697.

SLEPIAN, D. (1983), 'Some comments on Fourier analysis, uncertainty and modeling', *SIAM Review*, Vol. 25, No. 3, July, pp. 379–393.

STANLEY, W.D., DOUGHERTY, G.R. and DOUGHERTY, R. (1984), *Digital signal processing*, Reston Publishing Company Inc., Reston, Virginia, 2nd ed.

TAYLOR, L.S. (1981), 'The phase retrieval problem', *IEEE Transactions on Antennas and Propagation*, Vol. AP-29, No. 2, March, pp. 386–391.

THOMAS, B.MacA. (1976), 'Crosspolarisation characteristics of axially symmetric reflectors', *Electronics Letters*, Vol. 12, No. 9, April, pp. 218–219.

THOMAS, B.MacA. (1986), 'A review of the early developments of circular-aperture hybrid-mode corrugated horns', *IEEE Transactions on Antennas and Propagation*, Vol. AP-34, No. 7, July, pp. 930–935.

THOMPSON, A.R., MORAN, J.M. and SWENSON, JR., G.W. (1986), *Interferometry and synthesis in radio astronomy*, John Wiley & Sons, New York.

TRICOLES, G. and FARHAT, N.H. (1977), 'Microwave holography: applications and techniques', *Proceedings of the IEEE*, Vol. 65, No. 1, January, pp. 108–121.

UYTTENDAELE, A.G. (1986), 'Evolution of antenna sidelobe regulation', *IEEE Transactions on Broadcasting*, Vol. BC-32, No. 4, December, pp. 85–88.

WALTER, C.H. (1965), *Traveling wave antennas*, McGraw-Hill Electronic Science Series, McGraw-Hill Book Company, New York.

WESTCOTT, B.S. and BRICKELL, F. (1979), 'Dual offset reflectors shaped for zero cross-polarisation and prescribed aperture illumination', *Journal of Physics D: Applied Physics*, Vol. 12, No. 2, February, pp. 169–186.

WOOD, P.J. (1980), *Reflector antenna analysis and design*, IEE Electromagnetic Waves Series; 7, Peter Peregrinus Ltd, Stevenage, UK.

WOOD, P.J. (1987), 'A near field test system for very large antennas', in *Fifth International Conference on Antennas and Propagation (ICAP 87), Part 1: Antennas*, IEE Conference Publication Number 274, York, March, pp. 489–492.

YAGHJIAN, A.D. (1986), 'An overview of near-field antenna measurements', *IEEE Transactions on Antennas and Propagation*, Vol. AP-34, No. 1, January, pp. 30–45.

YEH, Y.-C. (1949), 'The received power of a receiving antenna and the criteria for its design', *Proceedings of the IRE*, Vol. 37, No. 2, February, pp. 155–158.

GLOSSARY

This glossary provides succinct definitions of abbreviations and mathematical notations that are used throughout the thesis. Fuller definitions are referred to by page number or by equation number.

$A^I(x,y)$	Alias of $f(x,y)$ (3.25)
$A_{\rm m}(u,v)$	Measured far field amplitude pattern (pp. 129, 147)
Alias$\{\cdot\}$	Aliasing operator (3.25)
D	Diameter of antenna aperture
$D^{A_{\rm m}}$	Diameter of $S^{A_{\rm m}}$ (p. 147)
DFT	Discrete Fourier transform (p. 93)
$E_{\rm c}$	Corrected envelope error (4.26)
$E_{\rm m}$	Measured envelope error (4.24)
$\mathbf{E}(\mathbf{r})$	Electric field vector
$E_x(\mathbf{r})$	x component of $\mathbf{E}(\mathbf{r})$ (1.5)
$E_{\rm co}(\mathbf{r})$	Copolar component of $\mathbf{E}(\mathbf{r})$ (p. 46)
$E_{\rm x}(\mathbf{r})$	Cross polar component of $\mathbf{E}(\mathbf{r})$ (p. 46)
$\mathbf{E}_{\rm i}(\mathbf{r})$	Incident electric field (p. 22)
$\mathbf{E}_{\rm r}(\mathbf{r})$	Reflected electric field (p. 22)
$\mathbf{E}_{\rm a}(x,y)$	Aperture field distribution (p. 50)
$\mathbf{E}_{\rm f}(\theta;\phi)$	Feed pattern (p. 50)
$\grave{\mathbf{E}}(u,v)$	Fourier Fresnel pattern (2.39)
$\acute{\mathbf{E}}(u,v)$	Far field pattern (2.29)
$\mathbf{e}_0$	Complex value of a plane wave field at the origin (2.1)
$\mathbf{e}(\mathbf{r})$	Geometric optics complex function (2.10)
f	Focal length of a paraboloid
$f(x,y)$	Copolar aperture field distribution (3.74) or image (p. 85)
$f^*(x,y)$	Complex conjugate of $f(x,y)$
$\tilde{f}(x,y)$	Conjugate reflection of $f(x,y)$ (3.39)
$f_{\rm a}(x,y)$	Actual copolar aperture field distribution (pp. 130, 140)
$f_{\rm c}(x,y)$	Corrected copolar aperture field distribution (p. 154)
$f_{\rm d}(x,y)$	Design copolar aperture field distribution (pp. 129, 135)
$f_{\rm e}(x,y)$	Estimated copolar aperture field distribution (p. 130)
$f_{sl}(x,y)$	Speaker aperture field distribution of speaker l (p. 237)
$f^{\rm x}(x,y)$	Cross polar aperture field distribution (p. 149)
$\lvert f(x,y)\rvert$	Amplitude of $f(x,y)$ (1.3)
$[f]$	Available information about $f(x,y)$ (p. 104)
$ff(x,y)$	Autocorrelation of $f(x,y)$ (p. 89)
$ff_{\rm m}(x,y)$	Estimate of $ff_{\rm a}(x,y)$ obtained from $A_{\rm m}(u,v)$ (p. 184)

$F(u,v)$	Copolar far field pattern (3.74) or Fourier transform (p. 85)
$f[m,n]$	Sampled image $f(x,y)$ (3.19)
$F[p,q]$	Sampled Fourier transform $F(u,v)$ (3.19)
$\mathcal{F}(\zeta,\gamma)$	Z-transform of $f(x,y)$ (3.48)
$\tilde{\mathcal{F}}(\zeta,\gamma)$	Z-transform of conjugate of $\tilde{f}(x,y)$
$\mathcal{F}\mathcal{F}(\zeta,\gamma)$	Z-transform of conjugate of $ff(x,y)$
FFT	Fast Fourier transform (p. 96)
FT$\{f(x,y)\}$	Fourier transform of $f(x,y)$ (p. 89)
$G(\theta;\phi)$	Gain pattern (1.21)
$G_{\max}$	Peak gain (1.2.2)
GO	Geometrical optics (p. 24)
GTD	Geometrical theory of diffraction (p. 26)
$\mathbf{H}(\mathbf{r})$	Magnetic field vector
IDFT	Inverse discrete Fourier transform (p. 95)
IFT$\{F(u,v)\}$	Inverse Fourier transform of $F(u,v)$ (p. 89)
$\hat{\imath}$	Polarization unit vector (1.17)
III(x,y)	Grid of delta functions (p. 89)
j	Imaginary unit ($j^2=-1.0$)
$\mathbf{J}(\mathbf{r})$	Electric current density (1.10)
$\mathbf{J}_s(\mathbf{r}')$	Surface electric current density (1.11)
$\mathbf{J}_m(\mathbf{r})$	Magnetic current density (1.10)
$\mathbf{J}_{ms}(\mathbf{r}')$	Surface magnetic current density (1.11)
k	Wave number (1.15)
L_x^f	Extent of $f(x,y)$ in the x direction (p. 86)
$\max(f(x,y))$	Maximum value of $f(x,y)$
$\hat{\mathbf{n}}$	Unit vector normal to a surface
$n(x,y)$	Pseudo random noise function (p. 142)
$p(\mathbf{r})$	Acoustic pressure (p. 222)
$\acute{p}(\mathbf{r})$	Acoustic pressure in far field (p. 222)
P	Power
$\mathbf{P}(\mathbf{r})$	Complex Poynting vector (1.16)
phase$\{f(x,y)\}$	Phase of $f(x,y)$ (1.3)
PO	Physical optics (p. 29)
r	Radial distance from an antenna
$\mathbf{r}$	Arbitrary position vector
$\hat{\mathbf{r}}$	Unit vector parallel to $\mathbf{r}$
$\mathbf{r}'$	Position on a surface
R_{ff}	Far field distance (1.20)
R	Radius of radiation hemisphere (p. 30)
$(r;\theta;\phi)$	Spherical coordinate (p. 3)
ran$(\cdot,\cdot)$	Random distribution (p. 142)
real$\{f(x,y)\}$	Real part of $f(x,y)$
rect(x,y)	Rectangle function (p. 89)
rms	Root mean square
$\hat{\mathbf{s}}_0$	Unit vector parallel to direction of propagation of a plane wave (2.1)
$s(\mathbf{r})$	Geometric optics phase function (2.10)
S	Surface
S^{aper}	Support of aperture (p. 140)

S^{auto}	Support of autocorrelation of aperture field
S^f	Support of $f(x,y)$ (p. 86)
$\text{Samp}\{\cdot\}$	Sampling operator (3.19)
$\text{sinc}(x,y)$	Sinc function (p. 89)
t	Time
T	Temperature
(u,v)	Far field (or Fourier transform) plane coordinate (2.28)
V	Volume
$w(u,v)$	A particular function (2.28)
$\hat{\mathbf{x}}$	Unit vector in the x direction
(x,y)	Aperture (or image) plane coordinate (p. 32)
(x,y,z)	Cartesian coordinate (1.7)
$Z^{\mathcal{F}}$	Zeros of $\mathcal{F}(\zeta,\gamma)$
α_x	Sampling factor in x direction (3.22)
Γ_{cal}	Model parameter characterizing calibration inaccuracy (p. 147)
Γ_{off}	Design envelope offset (p. 152)
Γ_{ran}	Model parameter characterizing far field measurement noise (p. 147)
$\delta(x,y)$	Delta function (p. 89)
Δ_x	Sample spacing in x direction (3.19)
$\Delta a(x,y)$	Aperture amplitude deviation (4.9)
$\Delta n(x,y)$	Reflector shape defect (p. 64)
Δx	Displacement of feed in x direction (p. 66)
$\Delta\psi(x,y)$	Aperture phase deviation (p. 64)
ϵ	Electric permittivity (1.10)
$\mathcal{E}^{\text{aa}}$	Aperture amplitude error (4.67)
$\bar{\mathcal{E}}^{\text{aa}}$	Aperture data error (4.68)
$\mathcal{E}^{\text{ap}}$	Aperture phase error (4.21)
$\mathcal{E}^{\text{ap}}_{\text{targ}}$	Target value for $\mathcal{E}^{\text{ap}}$ (p. 157)
$\mathcal{E}^{\text{auto}}$	Autocorrelation error (4.50)
$\mathcal{E}^{\text{fa}}$	Far field amplitude error (4.23)
$\bar{\mathcal{E}}^{\text{fa}}$	Far field data error (4.51)
$\mathcal{E}^{\text{F}}$	Fourier error (3.60)
$\mathcal{E}^{\text{I}}$	Image error (3.71)
η	Efficiency
$(\theta;\phi)$	Angular direction (p. 9)
$(\theta_{\text{az}};\theta_{\text{el}})$	Direction given in azimuth and elevation angles (p. 231)
λ	Wavelength (1.15)
$\Lambda_{\text{d}}(u,v)$	Design envelope (p. 152)
μ	Magnetic permeability (1.10)
ν	Position along a diagonal cut in the far field plane (4.65)
ξ	Position along a diagonal in the aperture plane (4.8)
ρ	Electric charge density (1.10) or normalized radius (3.6)
ρ_{m}	Magnetic charge density (1.10)
ρ_{ms}	Magnetic surface
σ	Conductivity (1.10)
τ_{quad}	Model parameter characterizing aperture amplitude taper (4.13)
τ_{ran}	Model parameter characterizing complex aperture noise (4.14)

ψ_{pan}	Model parameter characterizing amount of panel displacement (4.10)
ψ_{quad}	Model parameter characterizing defocus (4.10)
ψ_{sl}	Phase of signal feeding speaker l (p. 237)
ω	Angular frequency (1.1)
Ω_{pan}	Model parameter characterizing area of displaced panel (4.12)
$\odot$	Convolution operation (p. 89)

INDEX

Where special terms are first defined, they are also typed in italics. This index merely lists these terms with the page number on which they are defined.